UNIVERSITÀ DEGLI STUDI DI FERRARA
CORSO DI LAUREA IN INFORMATICA

# *Analisi di dati satellitari riguardanti la relazione tra vegetazione e infezioni malariche in Mozambico*

*Relatore:*
**Prof. Guido SCIAVICCO**

*Laureando:*
**Matteo RANGONI**

ANNO ACCADEMICO 2023 – 2024

# *Indice*

**Page**

*Capitolo 1*

# *Introduzione*

La malaria è una malattia infettiva che colpisce milioni di persone in tutto il mondo e continua a essere una delle principali sfide per la sanità globale. Il Mozambico è tra i Paesi più colpiti da essa, e gli sforzi per contrastarla sono ingenti e in continua evoluzione. La causa principale della trasmissione del parassita sono le zanzare che, come è noto, sono l'animale più letale al mondo. La tipologia di copertura del suolo e i suoi cambiamenti, come la deforestazione, possono influenzare la trasmissione della malaria e la distribuzione delle zanzare. Questo è importante perché suggerisce che la preservazione del territorio e delle foreste potrebbe rientrare tra le strategie di lotta alla malaria. La presente tesi si propone di analizzare la relazione tra vegetazione e infezioni malariche in Mozambico, utilizzando dati satellitari e informazioni epidemiologiche nel periodo dal 2000 al 2022. L'obiettivo è quello di comprendere come la copertura del suolo, in particolare la presenza di foreste, possa influenzare la diffusione della malaria e se la deforestazione possa contribuire ad un aumento dei casi. Le motivazioni che hanno portato ad approfondire questo tema sono molteplici. Inizialmente, lo studio è nato su richiesta di ricercatori nell'ambito economico, interessati all'estrazione di dati riguardanti la deforestazione e la malaria in Mozambico, nel periodo temporale più ampio comune nei dati. Inoltre, la relazione tra vegetazione e malaria, sebbene sia stata

confermata da diversi studi nelle regioni del Sud America e dell'Asia, non è ancora chiara per quanto riguarda l'Africa, dove si verificano la maggior parte dei casi. Infine, il Mozambico, in particolare, rientra tra i dieci Paesi maggiormente colpiti sia dalla malaria che dalla deforestazione a livello globale. Un altro obiettivo di questa tesi, considerando l'uso di dati satellitari, è l'analisi del territorio attorno a villaggi e miniere nel Mozambico, fornendo poi risultati a supporto di ulteriori studi socioeconomici. Le metodologie utilizzate sono principalmente di analisi dei dati, analisi spaziale, statistica e di immagini e l'uso di metodi di apprendimento automatico quali la regressione e il *clustering*. La tesi è strutturata come segue: dopo una presentazione dettagliata di tutte le fonti dei dati utilizzate nello studio, raccolte da diverse sorgenti, vi è, nel capitolo successivo, una descrizione delle caratteristiche di questi dati e del loro formato, con tutte le operazioni e analisi preliminari effettuate per la loro preparazione. Nel capitolo seguente, il quarto, viene esposta la metodologia di ricerca adottata e presentate le domande di ricerca che verranno affrontate nei capitoli successivi. Per ogni domanda, nei quattro capitoli seguenti, vengono esposte le tecniche utilizzate, i risultati ottenuti e la risposta alla domanda. Infine, nell'ultimo capitolo, traiamo le conclusioni unendo le risposte alle domande di ricerca ed evidenziando un importante relazione tra territorio e infezioni malariche.

*Capitolo 2*

# *Fonte dei Dati*

In questo capitolo verranno presentate le fonti dei dati utilizzati per lo studio che sono stati raccolti da diverse piattaforme e organizzazioni. Verranno descritti i dataset, le informazioni che contengono, come sono stati creati dalla fonte e come sono stati ottenuti.

## 2.1 Africa Knowledge Platform

La piattaforma Africa Knowledge Platform[1] è un sito web ufficiale dell'Unione Europea, come si può evincere dal dominio contenente "europa.eu". Il portale contiene informazioni e dati riguardanti gli aspetti sociali, economici, territoriali e ambientali dei paesi dell'Africa raccolti dal Centro comune di ricerca (Joint Research Centre - JRC), il servizio scientifico e di conoscenza della Commissione europea. Questi dati, strumenti e informazioni presenti nel sito sono sviluppati internamente dal centro di ricerca europeo oppure in collaborazione con gli organi partner africani e internazionali. L'obiettivo è quello di avere un singolo punto d'ingresso per tutti i dati sull'Africa certificati e supportati dall'Unione Europea (EU) e dal JRC in modo da approfondire la collaborazione con l'Africa.

[1] https://africa-knowledge-platform.ec.europa.eu (visitato il 29/08/2024).

Nel sito sono presenti una serie di *dataset* provenienti sia dalla Commissione europea sia da terze parti, suddivisi per argomento e per obiettivo di sviluppo sostenibile. Questi sono accessibili filtrando anche per nazione, nel nostro caso il Mozambico. Per ogni dataset è possibile: visualizzarne i metadati che contengono tutte le informazioni sui dati e come interpretarli; scaricare direttamente i dati; e visitare il sito del progetto fonte da cui vengono estratte le informazioni. Inoltre è possibile aggiungere uno o più dataset alla mappa geografica rappresentante l'Africa per visualizzarli e filtrarli, ad esempio per anno, e creare mappe multi tematiche con scala da locale a continentale. Per alcuni dataset è possibile anche eseguire delle statistiche a livello nazionale.

Noi siamo interessati ai dataset identificati nella piattaforma come: Forest Cover, Land Cover Change (1995-2015), Forest Loss, Change in Malaria Prevalence among Children (%) 2000 - 2015, Malaria rates among children age 2 to 10 e, successivamente, Average annual precipitation (1981-2017) e Annual precipitation variability (L-CV). Questi, eccetto quelli sulle precipitazioni, sono tutti dataset di terzi parti. Essi, e altre fonti utilizzate per rispondere alle nostre domande, verranno approfonditi di seguito.

## 2.2 Global Forest Change

Nello studio condotto da Hansen et al. [14] sono stati esaminati i dati *Landsat*[2] globali con una *risoluzione spaziale* di 30 metri per qualificare l'estensione, la perdita e l'incremento delle foreste negli anni originariamente dal 2000 al 2012.

Il laboratorio Global Land Analysis and Discovery (GLAD) dell'Università del Maryland, in collaborazione con Global Forest Watch (GFW), fornisce poi dati aggiornati annualmente sulla perdita di foreste su scala globale[3], sempre ricavati

[2]Il programma Landsat consiste in una serie di missioni satellitari di osservazione della Terra gestite congiuntamente dalla NASA e dal Servizio geologico degli Stati Uniti (USGS) (`https://landsat.gsfc.nasa.gov/`) (visitato il 03/09/2024).

[3]Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. High-Resolution Global Maps of 21st-Century Forest Cover Change. Science 342 (15 November): 850-53. Data available on-line from: `https://glad.earthengine.app/view/global-forest-change`.

dalle immagini della serie temporale Landsat.

I dataset forniti di nostro interesse sono:

- Copertura arborea per l'anno 2000 (*treecover2000*), definita come chiusura delle chiome per tutta la vegetazione di altezza superiore a 5 metri. Questa è codificata come percentuale per cella della griglia di output.

- Incremento della copertura forestale globale nel periodo 2000-2012 (*gain*), definito come l'inverso della perdita, ovvero un cambiamento da non foresta a foresta durante il periodo di studio (dal 2000 al 2012). Codificato come valore 1 (guadagno) e 0 (perdita).

- Anno dell'evento di perdita di copertura forestale (*lossyear*) cioè la perdita di foresta nel periodo 2000-2023, definita come un cambiamento da uno stato di foresta a uno di non foresta oppure un *stand-replacement disturbance*[4]. Codificato come 0 (nessuna perdita) o come un valore compreso nell'intervallo 1-23, che rappresenta rispettivamente l'anno 2001-2023 dove la perdita è stata rilevata principalmente.

Sono poi presenti altri dataset come il *datamask* che rappresenta le aree mappate e le immagini spettrali di riferimento Landsat 7 sia per il primo anno disponibile (2000) che per l'ultimo (*last*) cioè ad oggi il 2023.

Solo il *lossyear* e il *last* sono aggiornati annualmente mentre gli altri fanno riferimento allo studio originale che copriva fino al 2012.

## 2.3 Land Cover Change

Il Land Cover (LC) è uno dei progetti dell'ufficio per il clima dell'European Space Agency (ESA)[5] dedicato alla generazione dell'Essential Climate Variable (ECV) per la copertura del suolo. La copertura del suolo (LC) è definita come ciò che concretamente è presente nella superficie terrestre che può essere alberi, erba, acqua, edifici, etc.

---

[4]Eventi nelle foreste che creano ampie aree prive di dominanza arborea e ricche di risorse fisiche e biologiche, tra cui il lascito dell'ecosistema precedente alla perturbazione [43].

[5]https://climate.esa.int (visitato il 03/09/2024).

Il dataset di nostro interesse fra quelli pubblicati è quello della serie di mappe Multi-Resolution Land Characteristics (MRLC) dal 1992 in poi. Questo è stato generato fino al 2015 direttamente dall'ESA Climate Change Initiative (CCI)[6] mentre dal 2016 dal Copernicus Climate Change Service[7] della Commissione europea.

Ogni valore corrisponde a un'etichetta che ne identifica la tipologia di terreno basata sulla classificazione Land Cover Classification System (LCCS) dell'United Nations Food and Agriculture Organization (UN FAO). I valori e il loro significato sono consultabili nella legenda CCI[8]. Questa è formata da delle classi coerenti a livello globale dette anche di *livello 1*, e sono 22 con un codice associato a ogni decina (10, 20, 30, etc). Le mappe CCI-LC sono poi descritte anche da una legenda più dettagliata chiamata di *livello 2* o *regionale*. Questa utilizza informazioni a livello regionale, dove possibile, e raggiunge un livello di dettaglio più elevato. Contiene quindi un numero maggiore di identificatori, e di conseguenza di classi, utilizzando i valori non decimali (e.g. 11, 12, etc.) e non sono disponibili in tutto il mondo [6].

Inoltre sono presenti anche 4 flag qualitativi in riferimento alla serie per l'intero periodo temporale, quindi non specifici per ogni anno. Forniscono informazioni per (i) contrassegnare le aree che non è stato possibile classificare; (ii) identificazione tramite satellite con 6 valori; (iii) numero di osservazioni satellitari valide e (iv) numero di anni in cui si sono verificati cambiamenti nella classe di copertura del suolo dal 1992 (Change Count).

L'insieme delle mappe annuali deriva da un unica mappa LC di riferimento generata dai dati Envisat MEdium Resolution Imaging Spectrometer (MERIS) Full Resolution (FR) di 300 metri e Reduced Resolution (RR) di 1000 metri di

[6]Defourny, P., Lamarche, C., Brockmann, C., Boettcher, M., Bontemps, S., De Maet, T., Duveiller, G. L. Harper, K., Hartley A., Kirches, G., Moreau, I., Peylin, P., Ottlé, C., Radoux J., Van Bogaert, E., Ramoino, F., Albergel, C., and Arino, O.: Observed annual global land-use change from 1992 to 2020 three times more dynamic than reported by inventory-based statistics, in preparation, 2023.

[7]Copernicus Climate Change Service, Climate Data Store, (2019): Land cover classification gridded maps from 1992 to present derived from satellite observation. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.006f2c9a (Accessed on 04-09-2024).

[8]European Space Agency. (2020). CCI Land Cover Maps Legend. European Space Agency. https://maps.elie.ucl.ac.be/CCI/viewer/download/CCI-LC_Maps_Legend.pdf.

risoluzione[9] dal 2003 al 2012. Indipendentemente da questa base[10], i Land Cover Change (LCC) sono rilevati a 1000 metri sulla base delle serie temporali AVHRR[11] tra il 1992 e il 1999, delle serie temporali SPOT-VGT[12] tra il 1999 e il 2013, dei dati PROBA-V[13] per gli anni dal 2014 al 2019 e di Sentinel-3[14] per gli anni recenti (dal 2020 al 2022).

Quando sono disponibili le serie temporali MERIS FR o PROBA-V, i cambiamenti rilevati a 1000 metri vengono rimappati a 300 metri. L'ultimo passo che viene fatto è poi aggiornare per gli anni precedenti e successivi le mappe LC dei 10 anni per produrre le 31 mappe annuali di LC dal 1992 al 2022 [6].

Il modulo dei cambiamenti funziona indipendentemente dal modulo di classificazione. Per evitare falsi rilevamenti di cambiamenti, dovuti alla variabilità durante gli anni delle classificazioni, ogni cambiamento deve essere confermato per più di due anni consecutivi nella serie temporale della classificazione.

Il cambiamento del terreno (LCC) viene rilevato solo tra le classi raggruppate secondo le sei categorie della classificazione di terreno Intergovernmental Panel on Climate Change (IPCC) che sono: terreni coltivati, foreste, pascoli, zone umide, insediamenti, e altro (arbusteti, vegetazione rada, aree spoglie e acqua). Questo per evitare che vengano individuati cambiamenti falsati tra classi LCCS semanticamente vicine e per requisito degli utenti come spiegato nella guida [6].

La corrispondenza tra le categorie di terreno IPCC (considerate per il rilevamento dei cambiamenti) e la legenda LCCS utilizzata nelle mappe CCI-LC è definita come seguente:

[9] https://www.esa-landcover-cci.org (visitato il 04/09/2024).

[10] Sviluppata da l'Université catholique de Louvain (UCL).

[11] L'Advanced Very High Resolution Radiometer (AVHRR) acquisisce misure della temperatura della superficie terrestre e marina, della copertura nuvolosa, della copertura di neve e ghiaccio, dell'umidità del suolo e degli indici di vegetazione, https://www.earthdata.nasa.gov (visitato il 04/09/2024).

[12] Lo SPOT-VEGETATION Programme (SPOT-VGT) ha monitorato la superficie terrestre del nostro pianeta per 16 anni (1998-2014) su base giornaliera, https://spot-vegetation.com (visitato il 04/09/2024).

[13] PROBA-V, lanciato il 7 maggio 2013, è un satellite miniaturizzato dell'ESA con il compito di mappare ogni due giorni la copertura del suolo e la crescita della vegetazione sull'intero pianeta, https://earth.esa.int (visitato il 04/09/2024).

[14] Missione satellitare europea di osservazione della Terra, https://sentinels.copernicus.eu (visitato il 04/09/2024).

1. Agriculture

    - 10: Cropland, rainfed
        - 11: Herbaceous cover
        - 12: Tree or shrub cover
    - 20: Cropland, irrigated or post-flooding
    - 30: Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)
    - 40: Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)

2. Forest

    - 50: Tree cover, broadleaved, evergreen, closed to open (>15%)
    - 60: Tree cover, broadleaved, deciduous, closed to open (>15%)
        - 61: Tree cover, broadleaved, deciduous, closed (>40%)
        - 62: Tree cover, broadleaved, deciduous, open (15-40%)
    - 70: Tree cover, needleleaved, evergreen, closed to open (>15%)
        - 71: Tree cover, needleleaved, evergreen, closed (>40%)
        - 72: Tree cover, needleleaved, evergreen, open (15-40%)
    - 80: Tree cover, needleleaved, deciduous, closed to open (>15%)
        - 81: Tree cover, needleleaved, deciduous, closed (>40%)
        - 82: Tree cover, needleleaved, deciduous, open (15-40%)
    - 90: Tree cover, mixed leaf type (broadleaved and needleleaved)
    - 100: Mosaic tree and shrub (>50%) / herbaceous cover (<50%)
    - 160: Tree cover, flooded, fresh or brakish water
    - 170: Tree cover, flooded, saline water

3. Grassland

    - 110: Mosaic herbaceous cover (>50%) / tree and shrub (<50%)
    - 130: Grassland

4. Wetland
    - 180: Shrub or herbaceous cover, flooded, fresh/saline/brakish water
5. Settlement
    - 190: Urban areas
6. Other
    - 120: Shrubland
        - 121: Evergreen shrubland
        - 122: Deciduous shrubland
    - 140: Lichens and mosses
    - 150: Sparse vegetation (tree, shrub, herbaceous cover) (<15%)
        - 151: Sparse tree (<15%)
        - 152: Sparse shrub (<15%)
        - 153: Sparse herbaceous cover (<15%)
    - 200: Bare areas
        - 201: Consolidated bare areas
        - 202: Unconsolidated bare areas
    - 210: Water bodies

## 2.4 Malaria Atlas Project

Il più grande database al mondo sulla malaria è archiviato dal Malaria Atlas Project (MAP).[15] Il loro obiettivo principale è quello di sviluppare la scienza della cartografia nella malaria [18] creando mappe di endemicità[16] globali. Lo scopo è quello di facilitare l'individuazione delle popolazioni a rischio, la previsione del

[15] https://malariaatlas.org (visitato il 05/09/2024).

[16] L'endemia è una misura del livello di rischio della malaria in una popolazione umana e determina l'età media della prima esposizione, il tasso di sviluppo dell'immunità e, quindi, lo spettro clinico previsto della malattia [18, 42, 41].

carico di malattia[17] (*burden*), dove intervenire e la misurazione dei progressi.

Come viene spiegato da Simon I Hay e Robert W Snow [18] viene fatto un assemblaggio fra i dati *training* epidemiologici e i dati ambientali *predictor* usando una serie di tecniche di mappatura statistica per metterli in relazione.

Viene affermato che il dato sul Parasite Rate (PR) costituisce la maggior parte delle informazioni disponibili globali sull'endemicità della malaria. Il PR è la proporzione di popolazione campionata che viene confermata positiva ai parassiti della malaria, normalmente identificandoli dai vetrini di sangue [10].

Nello studio in [18] hanno adottato una classificazione classica dell'endemicità della malaria [25] per standardizzare la definizione di rischio a livello globale. Questa fa riferimento al PR nell'età compresa tra i due e i dieci anni come: *Hypoendemic* se sotto il 10% per la maggior parte dell'anno; *Mesoendemic* tra l'11% e il 50%; *Hyperendemic* costantemente sopra il 50% e *Holoendemic* sopra il 75% ma questo fa riferimento solo alla fascia d'età di un anno.

Questi dati sul PR, che sono i più abbondanti, vengono raccolti (ulteriori informazioni in [13]) dal MAP attraverso letteratura, database e collaborazioni con partner, e usati per generare superfici di endemicità e, quindi, la mappa. Questa mappatura si differenzia dalle precedenti per tre aspetti: in primo luogo (i) perché è globale e mira a condividere i dati e il database e renderlo di pubblico dominio; (ii) usa dei criteri rigorosi per i dati, vengono presi in considerazione solo campionamenti di popolazione casuali o completi condotti dopo il 1985, dove la specie del parassita e i gruppi di età sono definiti, e che coinvolgono più di 50 persone, per minimizzare l'errore di campionamento; (iii) raccoglie dati oltre che sul *Plasmodium falciparum* anche sul *Plasmodium vivax* che viene invece spesso trascurato [18].

Per creare la mappa, vengono assemblati diversi dati sui fattori che influenzano il rischio di infezione da malaria e le conseguenze della malattia. Questi fattori sono la distribuzione dei principali vettori della malaria, le *Anopheles*[18], e la frequenza dei disturbi ereditari dell'emoglobina.

[17]Impatto negativo di una malattia su una popolazione in termini di cattivo stato di salute, rischio di decesso, costo delle cure o altri indici accreditati. (https://www.efsa.europa.eu/it/glossary/burden-disease) (visitato il 09/09/2024).

[18]Zanzara anofele.

Le zanzare anofele, vettori della malattia, sono molto sensibili al clima. Questo può essere utilizzato per predire la distribuzione della malaria nel tempo e nello spazio. Per mappare la malattia servono quindi anche dati sul territorio come temperatura, altitudine, precipitazione ed estensione della vegetazione (Sezioni 2.2, 2.3). Questi vengono abbinati con i dati rilevati del PR, nella posizione corrispondente, per mettere in relazione l'endemicità della malaria e l'ambiente. Come già visto nei capitoli precedenti, anche in [18] fanno uso di satelliti di osservazione terrestre per derivare queste informazioni.

Il PR è però una misura meno diretta per la trasmissione della malaria dal punto di vista della prevalenza. Inoltre, per le aree ad alta endemicità, i campioni di PR sono spesso limitati ai bambini, mentre nelle aree a bassa endemicità i rilevamenti sono di solito estesi a tutti i gruppi di età. Il PR è pertanto disturbato dai fattori di età della popolazione campionata, dal suo stato immunitario e dalla parassitemia periferica [18, 40, 39].

Nello studio [18] è stato necessario, quindi, trasformare il PR in altre misure della "forza di infezione" della malaria quali il *APfEIR*[19], il numero di punture infettive pro capite, spesso espresso annualmente [17, 40], la capacità vettoriale $C$[20] o il numero di riproduzione di base $R_0$[21]. Queste misure sono più dirette e sono maggiormente correlate al ciclo di vita e alle dinamiche delle popolazioni di zanzare. Tuttavia, ci sono molte meno misurazioni [15, 17] rispetto al PR, per questo sono state generate a partire dai dati PR e la conversione viene descritta negli articoli [15, 17].

I dati PR vengono quindi standardizzati in base all'età e mappati su tutta la superficie globale distribuendoli spazialmente. Il codice sorgente dei modelli per eseguire queste conversioni e tecniche di mappatura è reso disponibile pubblicamente[22] e i dati sono accessibili in un'interfaccia R[23], *malariaAtlas*[24] descritta in [30].

[19] Tasso entomologico di inoculo.

[20] Valore atteso di umani infettati per umano infetto, al giorno, assumendo efficienza di trasmissione perfetta [38].

[21] Il valore atteso di esseri umani infetti per ogni essere umano infetto, o il numero di zanzare infette per ogni zanzara infetta [38].

[22] `https://github.com/malaria-atlas-project` (visitato il 10/09/2024).

[23] `https://www.r-project.org/` (visitato il 10/09/2024).

[24] `https://cran.r-project.org/package=malariaAtlas` (visitato il 10/09/2024).

Sempre in [18] viene affermato che questi cambiamenti ambientali inevitabili [16] e globali influenzeranno le popolazioni a rischio malaria. Viene fatto l'esempio dei cambiamenti del terreno (LCC, vedi Sezione 2.3) in [7] come la deforestazione [12], che è quello di cui tratta questa tesi, che potrebbero modificare le dinamiche delle popolazioni delle zanzare vettore. Altri fattori sono la crescita della popolazione, l'urbanizzazione [15], il cambiamento climatico [20] che influenzano le dinamica della popolazione umana e la pandemia di HIV/AIDS, la denutrizione, lo stato socioeconomico [20] che influenzano la capacità della popolazione di far fronte alla malaria.

Il sito[25] offre mappe globali annuali, dal 2000 al 2022[26], con *risoluzione spaziale* di 5 km. Contengono dati sul rischio e *burden* della malaria, gli interventi che vengono fatti per contrastarla e i fattori connessi.

Vi sono due specie di parassiti, *Plasmodium falciparum (Pf)*, di nostro interesse, e *Plasmodium vivax*, e diverse metriche per i dati su di essi. Per il *Pf* sono presenti:

- *Parasite Rate* (PR), discusso precedentemente in 2.4, stima della percentuale di bambini di età compresa tra i 2 e i 10 anni che presentano parassiti di *Plasmodium falciparum* rilevabili.

- *Incidence Rate*, stima dei nuovi casi diagnosticati di *Pf* per 1.000 abitanti, in un determinato anno.

- *Incidence Count*, stima dei nuovi casi diagnosticati di *Pf* in un determinato anno.

- *Reproductive Number* ($R_0$, Sezione 2.4)

- *Mortality Rate*, stima dei decessi per *Pf* per 100.000 abitanti in un determinato anno.

- *Mortality Count*, stima dei decessi causati da *Pf* durante un determinato anno.

[25] https://data.malariaatlas.org (visitato il 12/09/2024).
[26] Nuova versione 2024-06.

Sono inoltre forniti dati riguardo gli interventi che vengono fatti in contrasto alla malattia, come zanzariere da letto trattate con insetticidi (Insecticide-Treated Nets - ITNs), l'irrorazione residua indoor (Indoor Residual Spraying - IRS) ovvero insetticida di lunga durata spruzzato nelle case[27] e medicine anti-malaria.

Vengono forniti anche dati sui fattori connessi accennati in 2.4 come emopatie, accessibilità (tempi di viaggio e strade), dati sulle zanzare, sulla popolazione umana, malaria zoonotica e altre malattie.

Nel sito sono presenti anche statistiche, per anno e per tipologia di parassita, a livello nazionale con le metriche descritte precedentemente per $Pf$ e $Pv$ e a livello sub-nazionale.

## 2.5 Immagini satellitari

Come abbiamo visto precedentemente le immagini satellitari sono fondamentali per la creazione di mappe e per l'analisi di dati geografici. Queste immagini sono acquisite da satelliti in orbita attorno alla Terra e permettono di osservare la superficie terrestre da una prospettiva spaziale.

Nel nostro studio sono state utilizzate le immagini del programma Landsat, utilizzate anche dal Global Land Analysis and Discovery (GLAD) come visto in Sezione 2.2. Questo programma è il più longevo per quanto riguarda l'acquisizioni di immagini satellitari della Terra ed è gestito congiuntamente dalla National Aeronautics and Space Administration (NASA) e dal United States Geological Survey (USGS)[28]. Il programma è iniziato nel 1972 con il lancio del primo satellite (Landsat 1) e ha continuato con l'invio di nuovi satelliti fino ad oggi con il Landsat 9.

Noi usiamo le immagini del Landsat 7, il settimo di questi satelliti, lanciato il 15 aprile 1999 e ancora attivo, ma sospeso dal gennaio 2024. La durata iniziale della missione era di cinque anni, mentre ad oggi è ancora in corso da oltre venticinque anni, coprendo quindi il periodo del nostro studio ovvero dal 2000 al 2022. Il satellite orbita attorno alla Terra a un'altitudine di 705 km (697 km dal 6 aprile 2022) in un'orbita quasi polare, eliosincrona, ha una risoluzione spaziale di 30

[27] https://www.cdc.gov (visitato il 12/09/2024).
[28] https://www.usgs.gov/ (visitato il 31/10/2024).

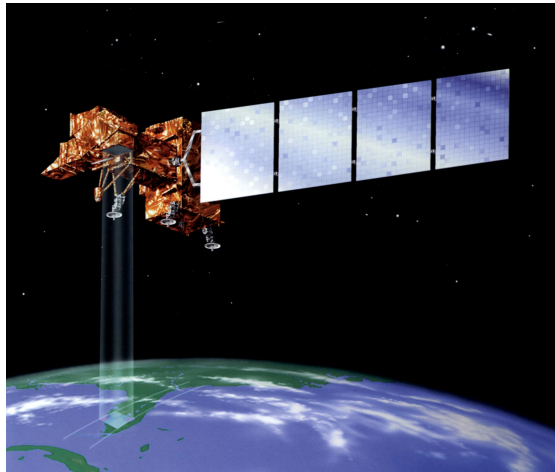



Figura 2.1: Illustrazione del satellite Landsat 7 in orbita (EROS, 2010).

metri e una risoluzione temporale di 16 giorni. Questo significa che ogni 16 giorni il satellite passa sopra la stessa area e acquisisce una nuova immagine [45].

Come descritto in [45] viene utilizzato il metodo di comunicazione *Direct downlink + SSR*. Direct downlink significa che solo quando sorvola le stazioni a terra, come ad esempio quelle di Sioux Falls (Dakota del Sud), Gilmore Creek (Alaska) e Svalbard (Norvegia), le immagini vengono trasmesse direttamente a terra con una velocità di 150 Mbps. Altrimenti, quando non è in vista di queste stazioni, vengono memorizzate nel *Solid State Recorder* (SSR) da 375 GB e trasmesse quando possibile.

Il sensore principale a bordo del Landsat 7 è l'Enhanced Thematic Mapper Plus (ETM+), un radiometro multispettrale a scansione, una versione migliorata degli strumenti Thematic Mapper di grande successo che erano a bordo di Landsat 4 e Landsat 5. Questo sensore è in grado di acquisire immagini in 8 bande spettrali, 7 visibili e una termica.

In questo studio sono stati utilizzati specificamente i dati *Landsat 7 ETM+ Collection 2 Tier 1 Surface Reflectance*. La *Collection 2* rappresenta la più recente rielaborazione dell'intero archivio Landsat, con significativi miglioramenti nella qualità dei dati. La *Tier 1* è la versione più accurata e completa dei dati, che soddisfa i requisiti di qualità geometrica e radiometrica.

Per quanto riguarda la *Surface Reflectance* (SR), è una misura della quantità di luce riflessa dalla superficie terrestre nelle diverse bande spettrali e il dataset

contiene la riflettanza superficiale e la temperatura della superficie terrestre corrette atmosfericamente. Questa misura è importante per gli studi ambientali perché permette di confrontare le caratteristiche della superficie indipendentemente dalle condizioni atmosferiche e dall'angolo del sole al momento della raccolta dei dati. A differenza dei dati normali TOA (Top of Atmosphere) non include l'interferenza atmosferica ed è quindi più adatta per confrontare le misurazioni effettuate in giorni diversi, a lungo termine o in aree con condizioni atmosferiche variabili. Inoltre è molto utile per calcolare gli indici di vegetazione come l'NDVI e l'EVI che vederemo in seguito.

Le immagini contengono quattro bande visibili e *near-infrared* (VNIR) e 2 bande *short-wave infrared* (SWIR) elaborate per ottenere una riflettanza superficiale ortorettificata e una banda *thermal infrared* (TIR) anch'essa elaborata per ottenere una temperatura superficiale ortorettificata. Contengono anche bande intermedie utilizzate per il calcolo dei prodotti ST (Surface Temperature) e bande QA (Quality Assessment).[29]

I dataset Landsat 7 SR sono creati con l'algoritmo Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) (versione 3.4.0). Inoltre i prodotti Collection 2 ST sono creati con un algoritmo single-channel sviluppato congiuntamente dal Rochester Institute of Technology (RIT) e dal NASA Jet Propulsion Laboratory (JPL).

I dati raccolti sono impacchettati in *tile* sovrapposte che coprono circa 170 km x 183 km utilizzando una griglia di riferimento standardizzata[30].

Inoltre, l'orbita del Landsat 7 è sincronizzata con il sole, cioè il satellite passa sopra la stessa area alla stessa ora solare locale in ogni passaggio. Ciò permette di confrontare le immagini acquisite in momenti diversi senza dover correggere la differenza di illuminazione. Si nota però che dal 2017 l'orbita si sta spostando verso un orario di acquisizione anticipato, con effetti dal 2019/2020 in poi [31].

Più precisamente le bande utilizzate disponibili nel USGS Landsat 7 Level 2, Collection 2, Tier 1 con SR corretta atmosfericamente (LANDSAT/LE07/C02/T1_L2) sono le seguenti:

[29] https://developers.google.com/earth-engine/datasets/catalog/LANDSAT_LE07_C02_T1_L2 (visitato il 31/10/2024).

[30] https://landsat.gsfc.nasa.gov/about/the-worldwide-reference-system/ (visitato il 31/10/2024).

- **SR_B1**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$
  - Offset: -0.2
  - Wavelength: 0.45-0.52 $\mu$m
  - Description: Band 1 (blue) surface reflectance
- **SR_B2**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$
  - Offset: -0.2
  - Wavelength: 0.52-0.60 $\mu$m
  - Description: Band 2 (green) surface reflectance
- **SR_B3**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$
  - Offset: -0.2
  - Wavelength: 0.63-0.69 $\mu$m
  - Description: Band 3 (red) surface reflectance
- **SR_B4**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$

- Offset: -0.2
  - Wavelength: 0.77-0.90 $\mu$m
  - Description: Band 4 (near infrared) surface reflectance

- **SR_B5**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$
  - Offset: -0.2
  - Wavelength: 1.55-1.75 $\mu$m
  - Description: Band 5 (shortwave infrared 1) surface reflectance

- **SR_B7**
  - Min: 1
  - Max: 65455
  - Scale: $2.75 \times 10^{-5}$
  - Offset: -0.2
  - Wavelength: 2.08-2.35 $\mu$m
  - Description: Band 7 (shortwave infrared 2) surface reflectance

## 2.6 Villaggi e miniere

Le informazioni sui villaggi provengono dal programma Demographic and Health Surveys (DHS), finanziato dall'United States Agency for International Development (USAID) che dal 1984 fornisce assistenza tecnica per più di 400 indagini in oltre 90 Paesi, migliorando la comprensione globale delle tendenze di salute e popolazione nei Paesi in via di sviluppo.[31]

Il programma DHS si è guadagnato una reputazione mondiale per la raccolta e la diffusione di dati accurati e rappresentativi a livello nazionale anche sulla malaria. Questi dati sono raccolti attraverso sondaggi a campione e sono disponibili per

[31] https://dhsprogram.com/ (visitato il 31/10/2024).

l'uso pubblico. Un'aspetto notevole è che la georeferenziazione avviene a livello di *cluster* piuttosto che a livello individuale. Questo comporta che ogni punto rappresenta un insieme di abitazioni e non singole case, garantendo così l'anonimato e la privacy degli individui, senza compromettere la riservatezza dei partecipanti alle indagini.

Più precisamente, vengono da noi utilizzati i punti centroidi dei villaggi, definiti come *cluster* di unità abitative (*household*) e i dati sulla malaria associati a essi. Questi dati sono: l'anno del test, il numero di individui testati, il numero di individui positivi e negativi al test e se c'è stato un intervento del governo contro la malaria con il periodo coperto da esso. Quest'ultimo, in particolare, è stato raccolto da più fonti.

La fonte dei dati sulle miniere è, invece, il sistema informativo PANGAEA [8], una biblioteca ad accesso libero finalizzata all'archiviazione, alla pubblicazione e alla distribuzione di dati georeferenziati provenienti dalla ricerca sul sistema terra. Il World Data Center PANGAEA è membro del World Data System (WDS) dell'International Science Council (ISC). Ospita inoltre il World Radiation Monitoring Center (WRMC) della Baseline Surface Radiation Network (BSRN) e come tale è accreditato come "Data Collection and Processing Center" (DCPC) del World Meteorological Organisation (WMO) Information System (WIS). [32]

I dati utilizzati sono le coordinate geografiche delle miniere all'interno della nazione ai quali è anche associata l'area di estensione della miniera stessa.

## 2.7 Precipitazioni

I dati per le precipitazioni sono stati ottenuti dal progetto Aquaknow[33] del Joint Research Centre della Commissione europea attraverso il sito Africa Knowledge Platform (Sezione 2.1). Questo progetto si occupa, dal 2002, di raccogliere e diffondere informazioni sulle risorse idriche nei Paesi in via di sviluppo, in particolare in Africa.

Vengono svolte ricerche sulla fattibilità, l'identificazione e la progettazione di soluzioni e sulla loro sostenibilità per migliorare la gestione della conoscenza per lo

[32] https://www.pangaea.de/ (visitato il 01/11/2024).
[33] https://aquaknow.jrc.ec.europa.eu/ (visitato il 01/11/2024).

sviluppo del settore idrico nei Paesi in via di sviluppo. L'obiettivo è di migliorare la raccolta di informazioni e dati, la comunicazione e la condivisione delle conoscenze tra le comunità di stakeholder coinvolte nello sviluppo del settore idrico [1].

Sono stati utilizzati i dataset *Average annual precipitation (1981-2017)* e *Annual precipitation variability (L-CV)* che coprono l'intera Africa. Il primo mostra le precipitazioni medie annuali (mm/anno) per tutto il periodo 1981-2017 in tutto il continente. Il secondo rappresenta la variabilità media delle precipitazioni (L-CV) intorno al valore medio annuale per lo stesso periodo. Quanto più grande è la L-CV, tanto più variabile è la precipitazione annuale da un anno all'altro.

*Capitolo 3*

# *Analisi Preliminare*

Iniziamo descrivendo come sono fatti i dati e le operazioni preliminari che sono state necessarie per l'elaborazione dei dataset.

## 3.1 Struttura dei Dati

Lo standard per i dati geografici è il formato Geographic Tagged Image File Format (GeoTIFF). Lo standard GeoTIFF è un'estensione del Tagged-Image File Format (TIFF) per la gestione di immagini *raster* georeferenziate o geocodificate. Il contenuto geografico supportato nei tag GeoTIFF comprende la proiezione cartografica, il *datum* (3.1.1), la dimensione del pixel al suolo e altre variabili geografiche, basate sul modello geodetico Epicentre 2.0 della società no-profit Petrotechnical Open Software Company (POSC) [34, 35].

I dati raster sono la controparte dei dati vettoriali, sono composti da una matrice di pixel, tutti di dimensione uguale. Sono formati da colonne verticali e righe orizzontali di pixel che contengono un valore. La dimensione, e di conseguenza il numero, di questi pixel determina la risoluzione.

Il formato TIFF è stato sviluppato nel 1986 da Aldus Corporation, che ha pubblicato l'ultima versione 6.0 nel 3 giugno 1992[1] e successivamente acquisita da Adobe[2]. È basato su tag e serve per l'archiviazione e lo scambio di immagini raster. TIFF fa da wrapper per diverse codifiche bitstream per immagini bitmap (raster).[3] La massima dimensione di un file TIFF è 4 GB.

### 3.1.1 Il formato GeoTIFF

GeoTIFF è stato creato nel 1995 da Niles Ritter. È nato dal fatto che il formato TIFF è limitato nelle applicazioni cartografiche, e non esisteva una struttura stabile e disponibile al pubblico per la trasmissione di informazioni geografiche [35]. Vi erano solo soluzioni private o limitate a specifiche applicazioni come il tag *geotie* di Intergraph.

Fornisce uno standard al pubblico per il supporto di immagini geografiche TIFF che possono provenire da satelliti, scansioni aeree, mappe o risultati di analisi geografiche. Utilizza i tag TIFF 6.0 per descrivere tutte le informazioni cartografiche associate alle immagini TIFF. Il suo scopo è quello di legare un'immagine raster a un modello di spazio o a una proiezione cartografica nota e di descrivere tali proiezioni [35]. Per immagazzinare le informazioni georeferenziate GeoTIFF utilizza solo 6 tag TIFF dedicati, in modo da non aver bisogno di farsi allocare nuovi tag da Aldus/Adobe. Per descrivere queste informazioni vengono usati codici numerici e un sistema chiave-valore chiamato GeoKey per codificarli.

Queste informazioni memorizzate nei tag sono i sistemi di coordinate. Ci sono tre spazi in cui sono definiti i sistemi di coordinate: (i) lo spazio raster, $R$, che fa riferimento ai valori dei pixel dell'immagine; lo spazio dei dispositivi (ii), $D$, che è legato al dispositivo fisico, ad esempio il monitor, e usa le coordinate $J$ e $I$ delle righe e delle colonne (rispettivamente) della matrice e non quelle $X$ e $Y$ geografiche; e (iii) lo spazio del modello, $M$, usato per fare riferimento ai punti della Terra [35].

---

[1] TIFF 6.0 Specifications, https://www.itu.int/itudoc/itu-t/com16/tiff-fx/docs/tiff6.pdf (visitato il 15/09/2024).

[2] https://www.adobe.com (visitato il 15/09/2024).

[3] https://www.loc.gov/preservation/digital/formats/fdd/fdd000022.shtml (visitato il 15/09/2024).

I dati raster sono costituiti da dati numerici spazialmente coerenti e memorizzati digitalmente. I valori dei dati raster sono organizzati in array bidimensionali, i cui indici sono utilizzati come coordinate. Lo spazio raster $R$ utilizza le coordinate I e J dello spazio $D$ che partono da 0,0 nell'angolo in alto a sinistra e aumentano con I a destra e J in basso.

Lo spazio del modello ($M$) riconosce le coordinate geografiche, geocentriche, proiettate e verticali. Per georeferenziare l'immagine è necessario specificare il sistema di coordinate utilizzato standard o, se non standard, definirlo. Le coordinate geografiche sono quelle che interessano i nostri dati.

I sistemi di coordinate geografiche mettono in relazione la *latitudine* e la *longitudine* angolare con un punto effettivo della Terra. La Terra viene rappresentata attraverso un ellissoide per approssimare la sua forma reale, geoide, che è molto complessa e in cartografia ne vengono utilizzati diversi tipi. La latitudine e longitudine sono gli assi delle coordinate del sistema di riferimento di punti su un ellissoide. La latitudine è definita come l'angolo sotteso al piano dell'equatore dell'ellissoide da una perpendicolare che attraversa la superficie dell'ellissoide a partire da un punto. La latitudine è positiva a nord dell'equatore e negativa a sud. La longitudine è definita come l'angolo misurato intorno all'asse minore (polare) dell'ellissoide da un meridiano primo (convenzionalmente Greenwich) al meridiano passante per un punto, positivo se a est del meridiano primo e negativo se a ovest [35].

Un'altra informazione è il datum geodetico, che viene identificato da un codice numerico, e serve affinché le coordinate descrivano in modo univoco un luogo mettendo in relazione la terra e l'ellissoide.

Le coordinate geografiche sono, quindi, univoche solo se identificate dal sistema a cui appartengono e se vengono forniti il meridiano primo e il datum geodetico.

Negli anni recenti, l'Open Geospatial Consortium (OGC)[4] ha pubblicato l'OGC GeoTIFF 1.1 standard (14 settembre 2019)[5] che formalizza la versione 1.0 delle specifiche e le allinea con la continua aggiunta di dati all'EPSG Geodetic Parameter Dataset. L'OGC GeoTIFF 1.1 è uno standard approvato dalla NASA Earth Science Data Systems e ampiamente utilizzato nei sistemi di dati della

[4] https://www.ogc.org (visitato il 15/09/2024).

[5] https://docs.ogc.org/is/19-008r4/19-008r4.html (visitato il 15/09/2024).

NASA Earth[6].

In sintesi, i dati raster sono formati da righe e colonne di pixel, e ogni pixel rappresenta un area geografica e il valore rappresenta un dato di quell'area. La dimensione fissa di questi pixel ne determina la risoluzione spaziale. Questi dati sono georeferenziati e quindi rappresentano un punto reale della superficie terrestre attraverso le coordinate. Queste immagini possono contenere anche più bande e quindi più valori per uno stesso pixel georeferenziato.

### 3.1.2 I nostri dati GeoTIFF

Ora possiamo parlare dei nostri dati in formato GeoTIFF. Vedremo come sono stati scaricati e come sono strutturati ovvero quante bande contengono, la risoluzione spaziale, il sistema di riferimento geografico, il tipo di dato, l'area e il periodo coperto.

Per i dati sulle foreste (vedi 2.2) la Terra è divisa in quadrati di 10x10 gradi con risoluzione spaziale di 1 arco secondo per pixel o, approssimativamente, 30 metri per pixel all'equatore. I dati contenuti sono valori a 8 bit senza segno. Per coprire l'intero Mozambico occorre scaricare le 3 "piastrelle" identificate dall'angolo in alto a sinistra dalle coordinate 20S, 30E; 10S, 30E; 10S, 40E.[7]

I dati per il cambiamento della tipologia di terreno, come spiegato in Sezione 2.3, si dividono in due fonti, la prima, dal 1992 al 2015, è l'ESA CCI. In questo caso è possibile scegliere se avere un file unico con 24 bande, una per ogni anno, oppure 24 file diversi, uno per ogni anno. Inoltre i file sono disponibili direttamente in formato GeoTIFF con solo la banda del LCC oppure in formato Network Common Data Form (NetCDF). Vengono scaricati attraverso il protocollo File Transfer Protocol (FTP). La seconda fonte, dal 2016 in poi, dal C3S CDS fornisce i dati in formato NetCDF che vanno trasformati in GeoTIFF utilizzando GDAL[8]. In entrambe le fonti, i file sono globali e con risoluzione spaziale di 300 metri.

[6] https://www.earthdata.nasa.gov/esdis/esco/standards-and-practices/geotiff (visitato il 15/09/2024).

[7] https://glad.earthengine.app/view/global-forest-change (visitato il 15/09/2024).

[8] GDAL è una libreria di traduzione per i formati di dati geospaziali raster e vettoriali, rilasciata sotto una licenza open source dalla Open Source Geospatial Foundation, https://gdal.org (visitato il 15/09/2024).

Per le mappe relative alla malaria (Sezione 2.4) è possibile scaricare i dati già ritagliati in una nazione specifica, nel nostro caso il Mozambico, oppure globali. Hanno risoluzione spaziale di 5x5 km e abbiamo sempre un file per ogni anno.

Nei dati riguardanti le precipitazioni (Sezione 2.7) abbiamo una mappa che rappresenta l'intero continente africano con risoluzione spaziale di circa 22x22 km per pixel. Dato che questi dati fanno riferimento all'intero periodo avremmo solo due file a banda singola, uno per la media annuale e uno per la variabilità annuale. Avendo bisogno di una certa precisione, questi file GeoTIFF hanno un tipo di dato Float32 ovvero un numero a virgola mobile a 32 bit, a differenza ad esempio di quelli del Land Cover, che, rappresentando numeri interi, sono a 8 bit.

Tutti i file GeoTIFF descritti precedentemente si utilizzano a banda singola, definita anche scala di grigi. Il sistema di riferimento geografico (usa latitudine e longitudine per le coordinate) è il WGS 84 (World Geodetic System 1984) identificato dall'EPSG:4326[9]. È basato su WGS 1984 ensemble (EPSG:6326) che ha una accuratezza limitata di 2 metri al massimo e ha riferimento dinamico (si basa su un datum che non è fissato sulla placca tettonica). È lo standard per il GPS e la navigazione aerea. Inoltre, la dimensione dei pixel non è mai esattamente quadrata, ma è presente una leggera differenza tra X e Y. Questo è comune nei dati geografici per compensare la curvatura terrestre.

Le immagini satellitari vengono estratte attraverso Google Earth Engine[10], una piattaforma basata sul cloud per l'analisi geospaziale su scala planetaria che mette a disposizione le enormi capacità di calcolo di Google per affrontare una serie di problemi sociali di grande impatto [11]. Earth Engine ospita immagini satellitari e le immagazzina in un archivio pubblico di dati che comprende immagini storiche della Terra risalenti a più di quarant'anni fa. Le immagini, aggiornate quotidianamente, sono poi rese disponibili per l'estrazione di dati su scala globale. Fornisce poi API[11] e altri strumenti per consentire l'analisi di grandi dataset. Noi usiamo le API di Earth Engine attraverso la libreria *ee* e il linguaggio Python[12] . Dopo aver eseguito l'autenticazione e la creazione del proprio progetto, è possibile inizializzare e utilizzare le funzioni di Earth Engine per estrarre i dati di interesse. Vengono

[9] https://epsg.io/4326 (visitato il 15/09/2024).

[10] https://earthengine.google.com/ (visitato il 02/11/2024).

[11] Application programming interface, un mezzo che consente ai componenti software di comunicare tra loro utilizzando i servizi Web.

[12] https://www.python.org/ (visitato il 02/11/2024).

fatte delle chiamate API gestite completamente dalla libreria che permettono di ottenere le immagini dal cloud e fare operazioni su di esse, utilizzando il Task Manager di Earth Engine. Si scelgono quindi le immagini di un'unico sensore, nel nostro caso, come spiegato in Sezione 2.5, il Landsat 7, che sono raggruppate in uniche "collection" di immagini. La libreria mette a disposizione una serie di operazioni, anche complesse, da fare sulle immagini che vedremo in seguito in Sezione 3.2. Questa serie di operazioni che scriviamo nel nostro *script* Python vengono assemblate in un DAG (Grafo Diretto Aciclico) che viene poi inviato a Earth Engine per la valutazione e l'esecuzione attraverso un modello di calcolo "pigro". I file GeoTIFF che si ottengono vengono esportati su Google Drive, e quelli della nostra specifica collezione (Landsat 7 Level 2, Collection 2, Tier 1) hanno una risoluzione spaziale minima di 30 metri per pixel per le bande multispettrali (visibili e infrarosse), 15 metri per la banda pancromatica e 60 metri per la banda termica, che viene però rilasciata con un *resampling* a 30 metri. Sono disponibili quindi, in formato GeoTIFF, le immagini di qualunque area del globo e di qualsiasi data in cui il satellite era attivo. Le bande del file GeoTIFF sono in formato 16 bit, e saranno tutte quelle disponibili elencate in Sezione 2.5, quindi non a banda singola come i precedenti. Il sistema di coordinate è sempre il WGS 84.

### 3.1.3 I dati tabellari

L'obiettivo sarà quello di trasformare questi dati raster in dati tabellari, in modo da poterli analizzare ed eseguire operazioni su di essi.

I dati riguardanti la malaria nei villaggi e la posizione delle miniere, però, sono stati forniti già pre-elaborati e in formato tabellare, più precisamente in formato CSV[13]. Ogni riga rappresenta un punto univoco e le colonne sono le *feature* di interesse. La posizione è rappresentata da due colonne per le coordinate geografiche, la latitudine e la longitudine, sempre in formato WGS 84 (EPSG:4326). Le altre colonne contengono l'identificativo e le informazioni di interesse spiegate precedentemente in Sezione 2.6.

[13]Comma-separated values, file di testo che contiene dati tabellari.

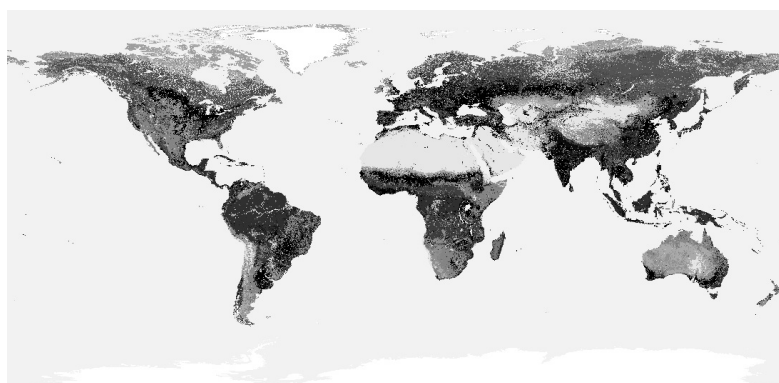



Figura 3.1: Mappa globale LC del 2022 (300x300 m).

## 3.2 Preparazione dei Dati

Per estrarre le informazioni di interesse dai dati provenienti dalle fonti descritte in precedenza, è necessario eseguire una manipolazione di questi dati.

I dati provenienti dalle fonti Global Forest Change (2.2), Land Cover (2.3) e Malaria MAP (2.4) sono in formato GeoTIFF e verranno eseguite delle operazioni di ritaglio e conversione della risoluzione, per poi trasformarli in dati tabellari per l'analisi. Saranno, inoltre, uniti in un unico dataset poiché ciò rappresenta uno degli obiettivi iniziali di questo studio.

I dati tabellari (villaggi e miniere, 2.6) sono già in formato CSV e verranno utilizzati direttamente senza bisogno di preparazione.

Infine sui dati satellitari (2.5) verranno eseguite altre operazioni, che vedremo in seguito, per poi essere ricondotti sempre a dati tabellari.

Possiamo visualizzare i nostri dati GeoTIFF nel software QGIS[14], come nell'esempio in Figura 3.1.

Questi file immagine GeoTIFF vengono decodificati in dataset attraverso il linguaggio Python e la libreria Xarray[15] e successivamente convertiti in dati tabellari, per la precisione pandas[16] *DataFrame*[17].

[14]https://www.qgis.org/ (visitato il 23/09/2024).
[15]https://xarray.dev/ (visitato il 23/09/2024).
[16]Pandas (https://pandas.pydata.org/ - visitato il 23/09/2024) è una libreria popolare

| | band | x | y | spatial_ref | band_data |
|---|---|---|---|---|---|
| 42351 | 1 | 34.687500 | -18.534854 | 0 | 130.000000 |

Tabella 3.1: Esempio di un punto del dataset LC del 2022 dopo essere stato convertito in dato tabellare con Xarray.

Come si può vedere in Tabella 3.1 sono presenti cinque *feature* ma solo tre sono di nostro interesse, la longitudine (x), la latitudine (y) e il valore (band_data).

La coppia di coordinate diventa quindi la nostra chiave che identifica ogni riga. La nostra tabella è quindi una tabella di coordinate con un valore associato a ogni punto. Questa è la forma più comune di rappresentazione dei dati raster.

### 3.2.1 Ritaglio

Per estrarre i dati riguardanti una specifica nazione è stato utilizzato uno *shapefile*, un formato di memorizzazione di dati vettoriali per caratteristiche geografiche. In particolare quello per i confini amministrativi del Mozambico[18] fornito dall'Ufficio per gli affari umanitari (*Office for the Coordination of Humanitarian Affairs - OCHA*) delle Nazioni Unite.

Il raster di dati Land Cover viene ritagliato utilizzando il ritaglio con maschera, impiegando come livello di maschera lo *shapefile* del Mozambico. Dalla mappa globale che rappresentava l'intero globo, si ottiene una mappa che rappresenta solo il Mozambico, ritagliata seguendo i confini della nazione. Questo processo è stato eseguito sempre con il software QGIS.

Per quanto riguarda i raster dei dati sulle foreste, le tre piastrelle che coprono l'intera nazione (20S, 30E; 10S, 30E; 10S, 40E) sono state prima unite (*merge*) e successivamente ritagliate con maschera.

Per i dati sulla malaria, invece, non è stato necessario il ritaglio perché, come spiegato precedentemente, la fonte offre i dati già estratti per nazione.

Anche i dati sulle precipitazioni sono stati ritagliati, dato che coprono l'intero continente africano, con lo stesso metodo.

open source di Python utilizzata per la manipolazione e l'analisi dei dati.

[17] Un DataFrame di pandas memorizza dati tabellari bidimensionali, mutabili per dimensione e potenzialmente eterogenei.

[18] https://data.humdata.org/dataset/cod-ab-moz (visitato il 24/09/2024).

### 3.2.2 Conversione della risoluzione

I dati hanno risoluzioni spaziali diverse: 30 metri per le foreste, 300 metri per Land Cover e 5 chilometri per la malaria. Questi vengono convertiti a 5x5 km per uniformarsi ai dati del MAP (malaria), che non subiscono né conversione né ritaglio.

Occorre quindi eseguire una diminuzione della risoluzione dei dati, che porterà alla trasformazione di un certo numero $n$ di pixel che componevano l'area (in questo caso di 5 km) in un'unico pixel. Questo comporta una riduzione del numero di pixel che costituiscono l'immagine e quindi a un numero minore di righe nei dati tabellari e, di conseguenza, a una perdita di informazioni (vedremo a breve come limitarla). Per comporre il pixel da 5 km, il numero di pixel dei dati riguardanti le foreste, che hanno risoluzione 30 metri, sarà quindi maggiore rispetto al numero di pixel utilizzati nei dati LC, che hanno risoluzione di 300 metri.

Questo cambiamento di risoluzione, che sintetizza più pixel in un solo pixel (nel nostro caso, ma è possibile anche il contrario) viene eseguito attraverso la trasformazione QGIS, che offre i seguenti metodi di ricampionamento:

- Vicino più Prossimo (*Nearest Neighbor*).
- Bilineare (Kernel 2x2): Media tra i 4 pixel al centro.
- Cubica (Kernel 4x4): Media tra i 16 pixel al centro.
- B-Spline Cubica (Kernel 4x4).
- Lanczos (Kernel 6x6).
- Media: ricampionamento medio, calcola la media ponderata di tutti i pixel che ci sono non *NODATA*.
- Moda: seleziona il valore che appare più spesso tra tutti i punti campionati. In caso di parità verrà selezionato il primo valore identificato come moda.
- Massimo: seleziona il valore massimo da tutti i pixel che contribuiscono non *NODATA*.
- Minimo: seleziona il valore minimo tra tutti i pixel che contribuiscono non *NODATA*.

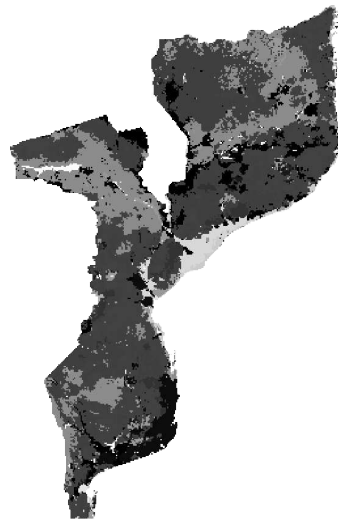



Figura 3.2: Mappa LC ritagliata al Mozambico e convertita con metodo moda del 2022 (5x5 km).

- Mediana: seleziona il valore medio di tutti i pixel che contribuiscono non *NODATA*.
- Primo Quartile (Q1)
- Terzo Quartile (Q3)

Il ricampionamento dei dati Land Cover è stato eseguito attraverso il metodo moda. Pertanto, l'area di 5 km, ora classificata con un determinato tipo di terreno, era costituita, precedentemente, per la maggior parte da quel determinato tipo di terreno. Trattandosi di un dato categorico non è possibile utilizzare altri metodi che eseguono operazioni aritmetiche sui dati, come la media, dato che si rischierebbe di introdurre nuovi valori che potrebbero non avere significato.

Questo approccio, che fa uso della moda statistica, limita la perdita di informazioni prendendo in considerazione tutte le celle (pixel) facenti parti dell'area originale a differenza dell'approccio predefinito *Nearest Neighbor*. Quest'ultimo, in sintesi, prende la cella al centro dell'area più vicino quindi considera solamente un pixel di 300 metri rispetto a tutti quelli presenti nell'area di 5 km, risultando impreciso.

Per la mappa *Tree Cover* che indica la percentuale di alberi presenti nell'area solamente per l'anno 2000, è stato utilizzato il metodo del ricampionamento medio.

Figura 3.3: Mappa *Tree Cover* ritagliata al Mozambico e convertita con metodo media del 2000 (5x5 km).

Questo effettua la media ponderata dei pixel nell'area ed è possibile giacché il dato è di tipo numerico. Rappresenta, inoltre, correttamente la percentuale di foresta che era presente nell'anno 2000 in quei 5x5 km, a differenza della media o del metodo predefinito.

Le restanti *feature* riguardanti gli alberi, *gain* e *loss*, definite in Sezione 2.2, sono trasformate con il metodo predefinito *Nearest Neighbor*. È stato utilizzato questo metodo a causa della frammentazione spaziale presente nei dati e della loro natura categorica. Utilizzando il metodo moda, infatti, si otterrebbe in entrambi i casi una mappa quasi interamente composta da valori pari a 0. Questo suggerisce che i cambiamenti o i guadagni rilevati con una risoluzione così alta (30 metri) risultano disomogenei e frammentati nel territorio. Tale metodo predefinito comporta, inoltre, una grande perdita di informazioni e imprecisione dei dati, dato che utilizza solo i 30 metri al centro dell'area di 5x5 km. Per questo motivo vedremo che tali dati non verranno usati per rispondere alle domande di ricerca. Questo vale anche per la colonna *Change Count* che traccia il numero dei cambiamenti del terreno in modo continuo, non limitandosi ad una cadenza annuale.

### 3.2.3 Trasformazione e unione

I dati riguardanti le foreste, il Land Cover e la malaria, una volta convertiti alla stessa risoluzione, vengono uniti in un unico dataset tabellare chiamato MOZ_Tree_Land_Malaria.

I file raster TIFF rimangono, anche dopo il ritaglio, delle matrici quadrate. I confini del Mozambico non hanno forma quadrata, quindi nelle matrici le celle al di fuori dai confini della nazione risultano nulle ($NaN$).

Quando si esegue la conversione in dataset di xarray e successivamente in DataFrame pandas, con il reset dell'indice, si ottiene una tabella che contiene le coordinate. A queste coordinate per i punti all'interno della nazione saranno associati dei valori nella colonna band_data mentre i punti esterni ai confini avranno valore nullo. Per ottenere nella tabella solo i punti all'interno della nazione non è possibile rimuovere semplicemente le righe con valori nulli. Questo perché altrimenti verrebbero rimossi anche i punti che non hanno valore ma sono all'interno dei confini. Tali punti sono necessari sia per l'unione fra i dataset in questione che per l'analisi dei dati.

Per ricavare una tabella con tutte le coordinate del Mozambico, utile per poi effettuare il *merge* con gli altri dataset, si fa uso dello *shapefile*[19]. Lo shapefile, che rappresenta in modo vettoriale i confini della nazione, viene rasterizzato. Si effettua una conversione da vettore a raster riempendo completamente l'interno dei confini di pixel con un valore *placeholder* (nel nostro caso la cifra "1") e della risoluzione adatta (5x5 km).

Salvando questo raster come file GeoTIFF, ed eseguendo uno script Python che lo trasforma in dataset tabellare, è ora possibile rimuovere le righe con valori nulli e ottenere la tabella con tutte e solo le coordinate di risoluzione 5x5 km della nazione. Vengono inoltre eliminate tutte le colonne che non servono, in questo caso anche quella dei valori della banda dato che sono tutti pari a "1", mantenendo solo le colonne $x$ e $y$ delle coordinate (vedi Tabella 3.1). Riordinando e rinominando le due colonne e azzerando l'indice si ottiene il DataFrame delle coordinate.

Grazie a queste coordinate che fanno da chiave avviene l'unione dei dataset in un'unico dataset tabellare attraverso la libreria pandas di Python. Ogni raster GeoTIFF viene decodificato in xarray dataset e da xarray dataset a pandas Da-

[19] https://data.humdata.org/dataset/cod-ab-moz (visitato il 24/09/2024).

taFrame. Le colonne *band* e *spatial_ref* (Tabella 3.1) che non sono necessarie, vengono rimosse. Le colonne vengono poi riordinate nell'ordine *y*, *x* e *band_data* e rinominate rispettivamente in *Latitude*, *Longitude* e con il nome della *feature* in questione. Successivamente avviene il *merge*, ovvero l'unione, della libreria pandas[20] simile al join di SQL[21] di tipo *inner*. La tipologia *inner* utilizza l'intersezione delle chiavi di entrambi i DataFrame, similmente a un SQL inner join, conservando l'ordine delle chiavi a sinistra e viene fatta sulle colonne *Latitude* e *Longitude* ovvero le coordinate che fungono da chiavi. Il primo *merge* viene effettuato sul DataFrame delle coordinate, ottenendo così il DataFrame finale. I successivi *merge* vengono eseguiti su quest'ultimo.

Per la colonna *loss year* vengono inoltre impostate tutte le celle con valore 23 a 0, escludendo quindi i dati riferiti all'anno 2023. Questo perché si prende in considerazione il periodo compreso tra il 2000 e il 2022.

Per i dataset composti da più file, uno per ogni anno, quindi copertura del suolo e malaria, viene eseguito un ciclo che scorre dal primo all'ultimo anno (da 2000 a 2022) che, cambiando sia il nome del file GeoTIFF di input che il nome della *feature*, li unisce al DataFrame finale.

Riguardo ai dataset contenenti i dati sulla malaria provenienti dal MAP, vengono anche rimosse tutte le righe che appartengono alla seconda banda (colonna *band* equivalente a 2). Queste righe sono in numero uguale a quelle della banda che contiene gli effettivi valori (come visto nella Sezione 3.1.1, i dati raster possono contenere più bande).

La colonna *Change Count*, che originariamente conterebbe i cambiamenti dall'anno 1992 all'ultimo anno disponibile, viene costruita per contare i cambiamenti nell'intervallo corretto (2000-2022). Questo si ottiene sottraendo i valori della colonna del 2022 a quelli della colonna del 2000 e unendo infine il risultato al DataFrame finale.

In Tabella 3.2 possiamo osservare un *data point* del dataset tabellare finale che è composto da 38653 righe e 98 colonne.

La latitudine e la longitudine sono univoche per ogni punto del dataset. Sono presenti tre *feature* per le foreste: copertura arborea dell'anno 2000, anno della

[20] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.merge.html (visitato il 28/09/2024).

[21] *Structured Query Language*, linguaggio standard per database.

| | Latitude | Longitude | Tree Cover % 2000 | Forest Loss Year | Forest Gain 2000-2012 | Land Cover 2000 | ... | Land Cover 2022 | Change Count | Incidence Rate 2000 | ... | Incidence Rate 2022 | Parasite Rate 2000 | ... | Parasite Rate 2022 | Incidence Count 2000 | ... | Incidence Count 2022 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -15.077847 | 30.229167 | 20.00 | 0.00 | 0.00 | 62.00 | ... | 62.00 | 0.0 | 0.338624 | ... | 0.282984 | 0.281792 | ... | 0.2363165 | 192.842 | ... | 236.438 |

Tabella 3.2: Esempio di un punto del dataset MOZ_Tree_Land_Malaria con tutte le *feature* incluse ($\mathbb{R}^{98}$), i dati dal 2000 al 2022 sono indicati con "..." per evitare ripetizioni.

perdita di foresta, guadagno di foresta dal 2000 al 2012, che non sono divise per anno. Sono incluse inoltre ventiquattro *feature* per la copertura del suolo, con la tipologia di copertura del suolo per ogni anno dal 2000 al 2022 e il numero dei cambiamenti totali. Infine, sono presenti sessantanove *feature* per la malaria divise in tre metriche ognuna per ogni anno dal 2000 al 2022.

Lo stesso processo di trasformazione in dato tabellare e *merge* è stato applicato ai dati sulle precipitazioni per unire le due metriche in un unico dataset tabellare con coordinate univoche.

### 3.2.4 Elaborazione dei dati satellitari

Per elaborare i dati satellitari si è utilizzato Google Earth Engine su Python parallelamente alla fase di ottenimento delle immagini. Le operazioni da eseguire, disponibili nella libreria, vengono definite prima di scaricare effettivamente le immagini. Queste operazioni, computazionalmente intensive, vengono poi eseguite sul cloud di Google tramite l'Earth Engine Task Manager per poi salvare le immagini già elaborate su Google Drive.

Per prima cosa viene definita l'area di studio, ovvero la geometria del Mozambico, attraverso il dataset FAO GAUL del 2015. Questo definirà l'area di lavoro per tutte le operazioni successive.

Viene poi eseguito il ciclo principale che scorre per ogni anno dal 2000 al 2022. Vengono definiti due intervalli di date, uno per la stagione principale dal 1 marzo al 25 luglio e uno per il periodo esteso che serve per il *gap-filling*[22] dal 1 febbraio al 25 agosto. Questi intervalli sono stati scelti per evitare la stagione delle piogge, dove sono presenti molte nuvole, e la stagione più secca, dove la vegetazione è più arida. Viene poi selezionata la collezione spiegata precedentemente, "LANDSAT/LE07/C02/T1_L2", e filtrata per l'area di studio e l'intervallo di date principale. Viene inoltre posto un filtro per la copertura di nuvole al 100%, perché avendo implementato un meccanismo che seleziona per ogni zona l'immagine con la copertura nuvolosa minore possibile nell'intervallo di tempo, non è necessario filtrarle ulteriormente alla base. Si scarica poi, allo stesso modo, la collezione per la data estesa e si procede ad applicare una serie di trasformazioni alle immagini.

[22] Riempimento dei dati mancanti nelle immagini.

Prima il mascheramento di nuvole e ombre eliminando i pixel identificati come nuvole e ombre di nuvole attraverso la banda *QA_PIXEL*.

Viene poi calcolato l'indice Normalized Difference Vegetation Index (NDVI) da aggiungere come banda. Questo indice, che varia tra -1 e 1, è una misura della salute della vegetazione. Un valore più alto indica una vegetazione più sana. Viene calcolata con la formula: $\text{NDVI} = \frac{NIR-RED}{NIR+RED}$ e viene aggiunta all'immagine.

Si calcola anche l'indice Enhanced Vegetation Index (EVI), che è una versione migliorata dell'indice NDVI che corregge alcuni problemi di saturazione e di rumore, con la formula:

$$\text{EVI} = G \times \frac{NIR - RED}{NIR + C_1 \times RED - C_2 \times BLUE + L}$$

Successivamente vengono convertite la bande in Float32 per una maggiore precisione e viene eseguita una funzione per ritagliare il confine con il Mozambico. Quest'ultima perché, nonostante la selezione dell'area di studio, le "piastrelle" ai confini che compongono le immagini scaricate possono contenere anche aree al di fuori del Mozambico.

Vengono fatte le stesse operazioni anche sulla collezione estesa e viene creata un'immagine composita dalla collezione di immagini satellitari precedentemente elaborate e filtrate. Questo attraverso una funzione mediana che calcola la mediana di tutti i pixel per ogni banda e per ogni pixel attraverso tutte le immagini della collezione.

Infine vengono riempiti gli eventuali *gap* dovuti a pixel nuvola rimossi o a immagini mancanti utilizzando la collezione estesa.

Vengono quindi selezionate le bande da esportare che sono: *SR_B1*, *SR_B2*, *SR_B3*, *SR_B4*, *SR_B5*, *SR_B7*, *NDVI*, *EVI* e viene chiamata la funzione per esportare l'immagine finale su Google Drive. Questa imposta la risoluzione a 5000 metri, il sistema di coordinate a EPSG:4326, il numero di pixel massimo a $1 \times 10^{13}$ e il formato a GeoTIFF.

Iterando questo per tutti gli anni avremmo immagini satellitari elaborate e chiare per il Mozambico dal 2000 al 2022 come, ad esempio, in Figura 3.4.

Per convertire queste immagini in dati tabellari si procede in modo differente rispetto ai dati precedenti. Si utilizza la libreria rasterio[23], una libreria Python

[23] https://rasterio.readthedocs.io/en/latest/ (visitato il 03/11/2024).

Figura 3.4: Immagine satellitare del Mozambico nel 2015 con risoluzione di 5x5 km, ottenuta dai dati Landsat 7, che mostra solo le bande visibili rosso (*SR_B3*), verde (*SR_B2*) e blu (*SR_B1*).

che consente di leggere e scrivere formati raster come GeoTIFF, fornendo un'API basata su array N-dimensionali di Numpy[24] e supporto per GeoJSON.

Iterando per ogni anno, si apre il file GeoTIFF con la libreria rasterio e si leggono tutte le bande del file in un array multidimensionale con il loro relativo nome. Si estraggono le coordinate e si creano due array unidimensionali per la latitudine e la longitudine, attraverso l'ausilio anche della libreria Numpy. Viene creato quindi il DataFrame pandas con le coordinate, l'anno in questione e i valori di ciascuna banda per ogni pixel. Tutti i DataFrame annuali vengono concatenati in un unico DataFrame (Tabella 3.3) che contiene tutti i dati tabellari con ogni riga rappresentante un pixel specifico con le sue coordinate, l'anno e i valori delle bande.

[24] https://numpy.org/ (visitato il 03/11/2024).

| Latitude | Longitude | Anno | SR_B1 | SR_B2 | SR_B3 | SR_B4 | SR_B5 | SR_B7 | NDVI | EVI |
|---|---|---|---|---|---|---|---|---|---|---|
| -10.487831 | 40.401730 | 2000 | 9011.5 | 9725.0 | 9374.5 | 18407.0 | 12311.0 | 9729.0 | 0.327690 | 3.194590 |
| -10.487831 | 40.446646 | 2000 | 8032.0 | 8677.0 | 8260.5 | 15976.5 | 9610.5 | 8211.0 | 0.317798 | 3.658406 |
| -10.487831 | 40.491561 | 2000 | 8421.0 | 8827.0 | 8573.0 | 13701.0 | 10163.0 | 8720.0 | 0.230224 | 4.435986 |
| -10.487831 | 40.536477 | 2000 | 8998.5 | 9488.5 | 9214.0 | 8941.0 | 8401.5 | 8080.5 | -0.007293 | 0.185292 |
| -10.532747 | 40.356814 | 2000 | 8250.0 | 9232.0 | 9198.0 | 15651.0 | 14047.0 | 10749.0 | 0.259689 | 1.799498 |

Tabella 3.3: Head del dataset dei dati satellitari MOZ_Sat con bande e indici vegetazionali.

## 3.3 Analisi Univariata

Passiamo ora all'analisi univariata, che considera una sola variabile alla volta. Ci concentriamo nel descrivere e osservare come cambia ogni singola caratteristica, senza considerare le altre. Questo tipo di analisi permette di ottenere una visione iniziale sulla pulizia e qualità dei dati, evidenziando eventuali anomalie o pattern. Inoltre, consente di fare ipotesi preliminari sul comportamento delle variabili, fornendo una base per analisi più complesse, come quelle bivariate o multivariate. Questa è stata eseguita su tutte le fonti.

### 3.3.1 Valori mancanti

Per iniziare è necessario analizzare il numero di celle nulle per ogni *feature* dei dataset. Questo ci permette di capire meglio come procedere con il preprocessing dei dati. I risultati di questa analisi preliminare riguardo al primo dataset creato precedentemente, per identificare le colonne con valori mancanti, sono riportati in Tabella 3.4.

Possiamo osservare che il numero di valori mancanti per colonna è relativamente basso rispetto alle 38653 righe totali del dataset. Tuttavia, si nota che il numero di pixel nulli per le *feature* riguardanti la malaria è maggiore rispetto alle *feature* di copertura del territorio e foreste, con solamente 3 righe nulle per il *Land Cover* e circa 552 e 175 per le *feature* di *Incidence* e *Parasite* rispettivamente. Questo potrebbe essere dovuto al fatto che i dati che descrivono le caratteristiche del territorio si estendono all'intera area della nazione, comprendendo anche zone disabitate come fiumi, laghi o montagne, dove i dati per la malaria potrebbero non

| **Colonna** | **Valori mancanti** |
|---|---|
| Tree Cover % 2000 | 3 |
| Forest Loss Year | 65 |
| Forest Gain 2000-2012 | 65 |
| Land Cover 2000 | 3 |
| ... | ... |
| Land Cover 2022 | 3 |
| Change Count | 141 |
| Incidence Rate 2000 | 552 |
| ... | ... |
| Incidence Rate 2015 | 553 |
| Incidence Rate 2016 | 552 |
| Incidence Rate 2017 | 552 |
| Incidence Rate 2018 | 552 |
| Incidence Rate 2019 | 552 |
| Incidence Rate 2020 | 552 |
| Incidence Rate 2021 | 556 |
| Incidence Rate 2022 | 552 |
| Parasite Rate 2000 | 175 |
| ... | ... |
| Parasite Rate 2022 | 175 |
| Incidence Count 2000 | 552 |
| ... | ... |
| Incidence Count 2015 | 553 |
| Incidence Count 2016 | 552 |
| Incidence Count 2017 | 552 |
| Incidence Count 2018 | 552 |
| Incidence Count 2019 | 552 |
| Incidence Count 2020 | 552 |
| Incidence Count 2021 | 556 |
| Incidence Count 2022 | 552 |

Tabella 3.4: Numero di valori nulli per ciascuna feature per il dataset MOZ_Tree_Land_Malaria. Le righe che presentato tutte lo stesso valore consecutivamente sono state omesse per brevità.

essere mappati. Inoltre, potrebbero esserci differenze nella metodologia di ritaglio dai confini della nazione utilizzata dal MAP.

In totale il numero di righe in cui manca almeno un valore è 600. Questo rappresenta circa lo 0,015% delle righe totali e, data la bassa percentuale, è stato scelto di rimuovere le righe contenenti valori nulli. Altri metodi per gestire la mancanza di dati possono includere l'imputazione tramite media o mediana, l'eliminazione della *feature* nel caso di una grande percentuale di valori mancanti o l'utilizzo di modelli predittivi per stimare i valori mancanti.

I dati sulle miniere non hanno valori mancanti e non è stata effettuata altra analisi preliminare su di essi, dato che, attraverso le colonne *Latitude* e *Longitude*, rappresentato semplicemente la posizione di ciascuna miniera. Anche per i dati pluviometrici non ci sono valori mancanti.

Nei dati riguardo ai villaggi e malaria è presente un punto *outlier*, cioè un valore che si discosta significativamente dagli altri, con coordinate vicine allo 0,0 che quindi viene rimosso. La colonna *gov_intervention* presenta 270 valori nulli. Di conseguenza vi sono 270 righe con almeno un valore nullo su 2023, ovvero circa il 13.3%. Questo significa che in 270 villaggi non è stata effettuata alcuna azione di intervento governativo contro la malaria.

Anche per quanto riguarda i dati satellitari non sono presenti valori mancanti, dato che, durante la conversione in dataset tabellare, sono state rimosse le righe con valori nulli. Questo per ottenere solo i pixel all'interno dei confini del Mozambico ed escludere i pixel nulli all'interno dell'area formati dalla mascheratura di nuvole e ombre o da immagini mancanti.

### 3.3.2 Grado informativo

Il grado informativo si riferisce alla quantità e qualità delle informazioni che possiede una colonna. Può essere descritto come l'utilità di una fonte di informazione e aiuta a determinare il valore dell'acquisizione di dati.

Questo viene calcolato in modo differente in base alla tipologia di dati presenti nella colonna, che possono essere categorici o numerici. I dati categorici, detti anche qualitativi, descrivono un'informazione che è rappresentata da un'etichetta e non sono misurati numericamente. Le categorie possono essere anche codificate

attraverso un numero, ma su di essi non hanno senso operazioni aritmetiche come la media. I dati categorici sono sempre discreti, quindi sono in numero finito, un esempio è la categoria di copertura del terreno, come erba o foresta, che nel nostro caso è codificata con un numero.

I dati numerici, invece, sono sempre rappresentati da numeri e descrivono delle misurazioni su cui hanno senso operazioni come la media. Questi possono essere anche continui, quindi in numero infinito, e offrono informazioni quantitative che è possibile elaborare matematicamente.

Per conoscere il grado informativo delle colonne numeriche è stata calcolata la varianza. La varianza descrive quanto i valori sono concentrati rispetto alla media e la sua unità di misura è il quadrato dell'unità del dato osservato. La formula per ottenerla è la seguente:

$$\text{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

La varianza è stata calcolata con il metodo *pandas.DataFrame.var* che restituisce la varianza *unbiased* ovvero normalizzata per $n - 1$.

Possiamo utilizzare la deviazione standard per visualizzare e commentare i risultati, che ha la stessa unità di misura del dato, ed è la radice quadrata della varianza.

| **Colonna** | **Deviazione standard** |
|---|---|
| Tree Cover % 2000 | 13.6143 |
| Change Count | 0.2436 |

Tabella 3.5: Deviazione standard di Tree Cover e Change Count.

La deviazione standard delle colonne numeriche non inerenti alla malaria nel dataset MOZ_Tree_Land_Malaria è mostrata nella tabella 3.5. La percentuale di foresta nell'anno 2000 nelle diverse aree del Mozambico varia tipicamente del 13.61% rispetto alla media.

La deviazione standard e la varianza delle colonne relative alla malaria (Tabella 3.6) sono state calcolate per ciascun anno. Questi parametri mostrano una notevole variabilità interannuale, indicando che la distribuzione dei dati cambia si-

| Colonna | Deviazione standard |
|---|---|
| Incidence Rate 2000 | 0.0865 |
| Incidence Rate 2001 | 0.0887 |
| Incidence Rate 2002 | 0.0864 |
| Incidence Rate 2003 | 0.0824 |
| Incidence Rate 2004 | 0.0835 |
| Incidence Rate 2005 | 0.0896 |
| Incidence Rate 2006 | 0.1014 |
| Incidence Rate 2007 | 0.1139 |
| Incidence Rate 2008 | 0.1185 |
| Incidence Rate 2009 | 0.1218 |
| Incidence Rate 2010 | 0.1298 |
| Incidence Rate 2011 | 0.1363 |
| Incidence Rate 2012 | 0.1399 |
| Incidence Rate 2013 | 0.1409 |
| Incidence Rate 2014 | 0.1451 |
| Incidence Rate 2015 | 0.1493 |
| Incidence Rate 2016 | 0.1466 |
| Incidence Rate 2017 | 0.1375 |
| Incidence Rate 2018 | 0.1242 |
| Incidence Rate 2019 | 0.1044 |
| Incidence Rate 2020 | 0.0848 |
| Incidence Rate 2021 | 0.0547 |
| Incidence Rate 2022 | 0.0459 |
| Parasite Rate 2000 | 0.1388 |
| Parasite Rate 2001 | 0.1492 |
| Parasite Rate 2002 | 0.1450 |
| Parasite Rate 2003 | 0.1339 |
| Parasite Rate 2004 | 0.1299 |
| Parasite Rate 2005 | 0.1299 |
| Parasite Rate 2006 | 0.1421 |
| Parasite Rate 2007 | 0.1566 |
| Parasite Rate 2008 | 0.1635 |
| Parasite Rate 2009 | 0.1659 |
| Parasite Rate 2010 | 0.1770 |
| Parasite Rate 2011 | 0.1859 |
| Parasite Rate 2012 | 0.1953 |
| Parasite Rate 2013 | 0.2041 |
| Parasite Rate 2014 | 0.2105 |
| Parasite Rate 2015 | 0.2092 |
| Parasite Rate 2016 | 0.1983 |
| Parasite Rate 2017 | 0.1775 |
| Parasite Rate 2018 | 0.1560 |
| Parasite Rate 2019 | 0.1330 |
| Parasite Rate 2020 | 0.1094 |
| Parasite Rate 2021 | 0.0761 |
| Parasite Rate 2022 | 0.0653 |
| Incidence Count 2000 | 1349.43 |
| Incidence Count 2001 | 1352.07 |
| Incidence Count 2002 | 1327.12 |
| Incidence Count 2003 | 1161.53 |
| Incidence Count 2004 | 925.45 |
| Incidence Count 2005 | 780.71 |
| Incidence Count 2006 | 691.58 |
| Incidence Count 2007 | 639.73 |
| Incidence Count 2008 | 639.87 |
| Incidence Count 2009 | 631.17 |
| Incidence Count 2010 | 713.98 |
| Incidence Count 2011 | 757.13 |
| Incidence Count 2012 | 822.12 |
| Incidence Count 2013 | 926.37 |
| Incidence Count 2014 | 898.78 |
| Incidence Count 2015 | 816.23 |
| Incidence Count 2016 | 757.07 |
| Incidence Count 2017 | 691.74 |
| Incidence Count 2018 | 711.20 |
| Incidence Count 2019 | 799.60 |
| Incidence Count 2020 | 916.29 |
| Incidence Count 2021 | 1070.39 |
| Incidence Count 2022 | 1208.73 |

Tabella 3.6: Deviazione standard di Incidence Rate, Parasite Rate, e Incidence Count (2000-2022).

| Colonna | Deviazione standard |
|---|---|
| tot_tested | 7.0387 |
| test_positive_malaria | 4.4118 |
| test_negative_malaria | 5.3485 |

Tabella 3.7: Deviazione standard delle colonne malariche nei villaggi.

| Colonna | Deviazione standard |
|---|---|
| SR_B1 | 1843.0215 |
| SR_B2 | 1604.8498 |
| SR_B3 | 1821.8140 |
| SR_B4 | 2131.2791 |
| SR_B5 | 1829.5866 |
| SR_B7 | 1412.6437 |
| NDVI | 0.0780 |
| EVI | 28.6626 |

Tabella 3.8: Deviazione standard delle bande satellitari e degli indici vegetazionali.

gnificativamente da un anno all'altro. I primi due insiemi di colonne rappresentano indici normalizzati su una scala da 0 a 1.

La deviazione standard per i sondaggi sulla malaria effettuati nei villaggi è presentata in Tabella 3.7. Questa fa riferimento a valori interi positivi che indicano un conteggio rilevato nei sondaggi.

Il grado informativo dei dati satellitari è espresso in Tabella 3.8. Possiamo notare che la banda NDVI, indice di salute della vegetazione, che va da -1 a 1, si discosta di 0.078 dal valore medio durante gli anni.

Per le colonne categoriche viene utilizzata l'entropia della colonna. L'entropia misura quanto è disordinata o incerta la distribuzione delle categorie nella colonna. Quando tutti gli elementi di una colonna appartengono a una singola categoria, l'entropia è minima perché è ordinata e c'è poca incertezza. Mentre quando, ad esempio, tutti gli elementi sono equamente distribuiti tra tutte le categorie l'entropia è massima e c'è molta incertezza. Si calcola come il valore atteso dell'autoinformazione, o "sorpresa", che è il logaritmo dell'inverso della probabilità

che l'evento accada [37]. Per calcolarla si fa riferimento alla seguente formula:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

Dove $n$ è il numero totale di categorie e $p(x_i)$ è la probabilità che l'evento accada, ovvero la probabilità di osservare la categoria $x_i$ nella distribuzione. La scelta della base del logaritmo ne determina l'unità di misura e varia a seconda delle applicazioni. Nel nostro caso viene calcolata attraverso il metodo *scipy.stats.entropy* della libreria SciPy[25] che permette di selezionare la base dell'algoritmo. È stata scelta la base $e$, valore default, che la restituisce in *nat*, ovvero l'unità naturale.

| **Colonna** | **Entropia** |
|---|---|
| Forest Loss Year | 0.4461 |
| Forest Gain 2000-2012 | 0.0124 |

Tabella 3.9: Entropia delle colonne Forest Loss e Forest Gain.

L'entropia delle *feature* categoriche riguardanti le foreste, ovvero l'anno di perdita e il guadagno nel periodo 2000-2012, è mostrata in Tabella 3.9.

Le categorie dell'anno di perdita sono gli anni dal 2000 al 2022, codificati con le ultime due cifre, e il valore 0 per nessuna perdita. In totale sono quindi 23 e, di conseguenza, il valore massimo dell'entropia sarà $\ln(23) = 3.1355$ *nat*. Le categorie del *Forest Gain* sono "guadagno" e "non guadagno", codificate rispettivamente con un'etichetta binaria 1 e 0. Sono quindi 2 e l'entropia massima sarà circa 0.6931 ($\ln(2)$) *nat*. I risultati in Tabella 3.9 si avvicinano al valore minimo, cioè 0. Ne vedremo la motivazione successivamente, osservando la distribuzione dei valori graficamente.

L'entropia delle restanti colonne categoriche, ovvero le tipologie di copertura del territorio, *Land Cover*, dall'anno 2000 all'anno 2022 è presentata in Tabella 3.10.

Le categorie sono rappresentante dalla tipologia di terreno e codificate con un numero, come visto in sezione 2.3. Il numero di categorie di terreno presenti in Mozambico per ogni anno varia da 20 a 22. L'entropia massima, nel caso fossero tutte distribuite equamente nel territorio, è quindi di circa 3 *nat*.

[25] https://scipy.org/ (visitato il 13/10/2024).

| **Colonna** | **Entropia** |
|---|---|
| Land Cover 2000 | 1.9016 |
| Land Cover 2001 | 1.9080 |
| Land Cover 2002 | 1.9108 |
| Land Cover 2003 | 1.9113 |
| Land Cover 2004 | 1.9117 |
| Land Cover 2005 | 1.9117 |
| Land Cover 2006 | 1.9121 |
| Land Cover 2007 | 1.9115 |
| Land Cover 2008 | 1.9120 |
| Land Cover 2009 | 1.9121 |
| Land Cover 2010 | 1.9121 |
| Land Cover 2011 | 1.9123 |
| Land Cover 2012 | 1.9130 |
| Land Cover 2013 | 1.9123 |
| Land Cover 2014 | 1.9111 |
| Land Cover 2015 | 1.9113 |
| Land Cover 2016 | 1.9106 |
| Land Cover 2017 | 1.9107 |
| Land Cover 2018 | 1.9121 |
| Land Cover 2019 | 1.9139 |
| Land Cover 2020 | 1.9149 |
| Land Cover 2021 | 1.9172 |
| Land Cover 2022 | 1.9209 |

Tabella 3.10: Entropia delle colonne Land Cover (2000-2022).

| **Colonna** | **Entropia** |
|---|---|
| year_test | 1.5199 |
| gov_intervention | 1.5753 |

Tabella 3.11: Entropia delle colonne year_test e gov_intervention.

Possiamo osservare l'entropia delle colonne categoriche relative ai villaggi e malaria, ovvero l'anno di test e l'intervento governativo, in Tabella 3.11.

Possiamo inoltre considerare l'entropia e, successivamente la distribuzione dei valori, anche per i dati satellitari prendendo in considerazione la colonna relativa all'anno. Questo perché non è presente lo stesso numero di punti per ogni anno e quindi può essere utile capire se ci sono anni che hanno grandi mancanze di punti, come era accaduto durante la fase di estrazione delle immagini satellitari. Il valore di entropia per la colonna Year del dataset MOZ_Sat è quindi di 3.1354 *nat*.

La distribuzione delle categorie nelle colonne, che ne determina l'entropia, è descritta nella sezione seguente.

### 3.3.3 Distribuzione dei valori

La distribuzione dei valori descrive come sono suddivise le categorie all'interno delle colonne, indicando quale categoria è presente maggiormente, quale lo è meno o se sono disposte equamente. Questo aiuta a comprendere meglio l'entropia, ed è applicabile solo alle colonne categoriche dato che contengono un numero finito di possibili valori. Date le $n$ categorie che formano la colonna viene calcolata la percentuale di occorrenze di ogni categoria per ogni *feature*.

Questo viene eseguito selezionando la colonna dal dataset e applicando la funzione *pandas.Series.value_counts* che restituisce il conteggio dei valori unici della colonna (frequenza assoluta[26]) con il parametro *normalize* impostato a *True* per ottenere la frequenza relativa[27].

[26] la frequenza assoluta di un valore $x$ è il numero di individui per cui la variabile assume tale valore $f_a(x) = |\{i : x_i = x\}|$.

[27] la frequenza assoluta rapportata al numero totale di individui del campione $f_r(x) = \frac{f_a(x)}{n}$.

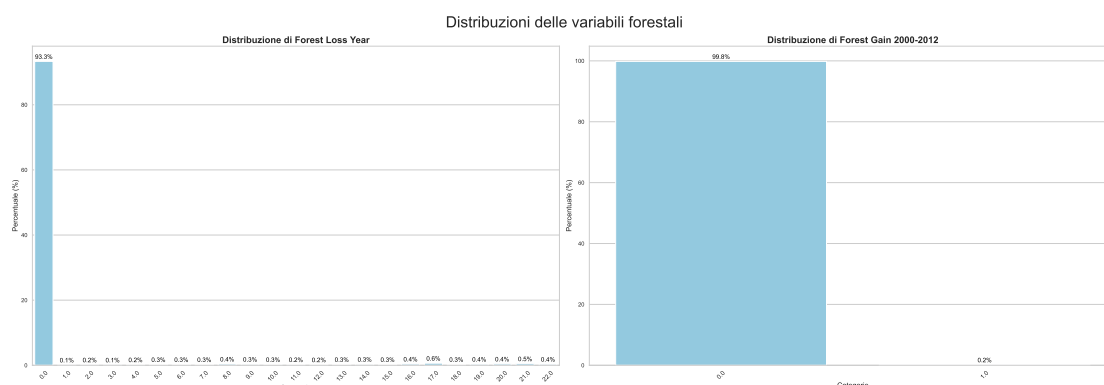



Figura 3.5: Grafici a barre della distribuzione delle *feature* forestali.

Il risultato è quindi un valore normalizzato tra 0 e 1 che viene moltiplicato per 100 per ottenere la frequenza percentuale che è la distribuzione dei valori mostrata nei grafici a barre.

In Figura 3.5 sono presenti le distribuzioni delle colonne categoriche riguardanti le foreste. Si può osservare che la maggior parte delle aree non ha subito perdite di foresta negli anni, come indicato dal valore 0 presente in percentuale del 93.3%. Questo spiega il motivo per cui l'entropia è bassa, come mostrato in Tabella 3.9. Lo stesso vale per il guadagno di foresta dove la maggior parte delle aree non ha avuto guadagni, con una percentuale del 99.8%.

Se vogliamo escludere il valore 0, che rappresenta l'assenza di perdita di foresta, possiamo osservare la distribuzione delle perdite di foresta negli anni nel grafico in Figura 3.6.

In Figura 3.7 sono presenti le distribuzioni delle colonne categoriche riguardanti la copertura del territorio. Si può osservare che la categoria prevalente è rappresentata dal numero 62 che corrisponde alla foresta latifoglie con alberi decidui con copertura della cella dal 15 al 40% e copre circa il 30% del territorio. Poi viene la categoria 60 che rappresenta la foresta latifoglie con alberi decidui (copertura maggiore del 15%) con il 25% del territorio e la categoria 120 che rappresenta l'arbusteto con il 20% del territorio. Si può notare, inoltre, la variazione di percentuale delle categorie di anno in anno, come per la foresta (62) che diminuisce nel tempo, o le aree urbane (190) che aumentano. La maggiore varietà nella distribuzione di queste categorie rispetto alle colonne forestali è confermata anche dall'entropia più alta.

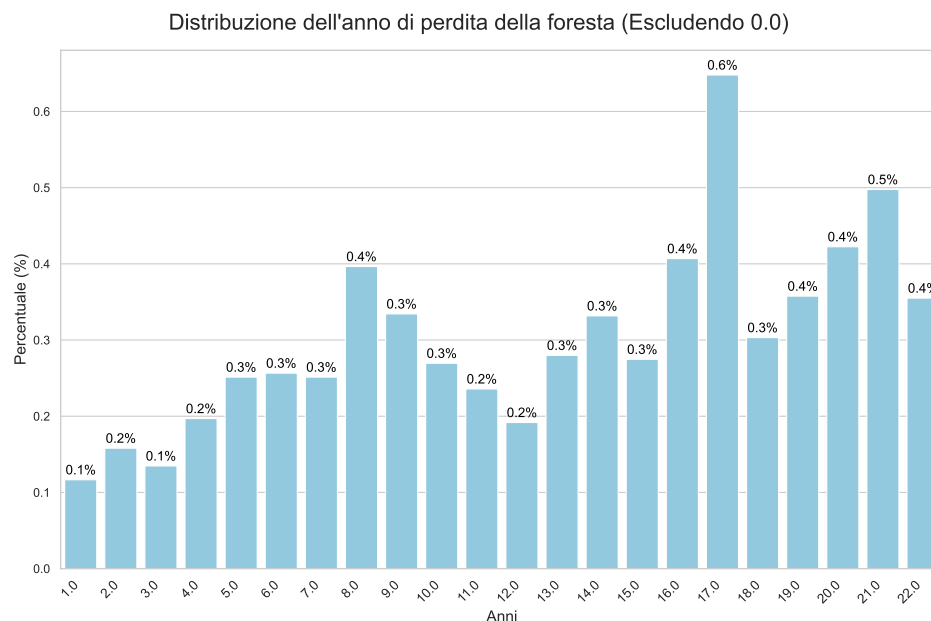



Figura 3.6: Grafico a barre della distribuzione della colonna *Forest Loss Year* escludendo il valore 0.0 (*no loss*).

A causa della bassa entropia, dello sbilanciamento nelle categorie e dell'imprecisione nel metodo di conversione dei dati, è stato scelto di escludere le colonne *Forest Loss*, *Forest Gain* e *Change Count*. Queste non offrivano informazioni significative all'interno del dataset ed erano di difficile interpretazione e legame con altre colonne per le analisi future.

La distribuzione delle colonne categoriche relative ai villaggi e alla malaria è mostrata in Figura 3.8. Si può osservare, per quanto riguarda l'anno di test, che la maggior parte dei test è stata effettuata nel 2011 e nel 2023. Per i trattamenti contro la malaria si può osservare che quasi la metà dei trattamenti, 44.6%, è stata effettuata dal 2015 al 2023.

Infine possiamo osservare la distribuzione dei punti negli anni nei dati satellitari, come vediamo in Figura 3.9. Gli anni con meno punti sono il 2004, il 2006 e il 2010.

Figura 3.7: Grafici a barre della distribuzione delle *feature* di copertura del suolo.

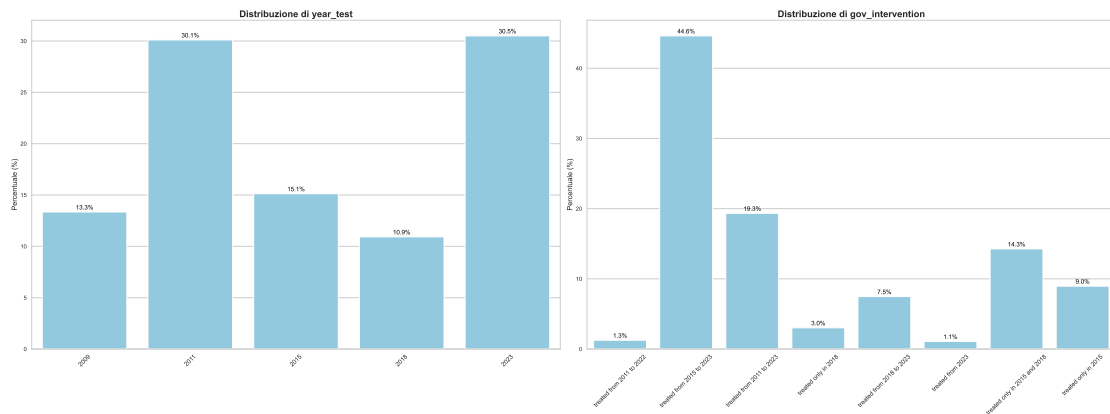



Figura 3.8: Grafici a barre della distribuzione delle colonne year_test e gov_intervention.

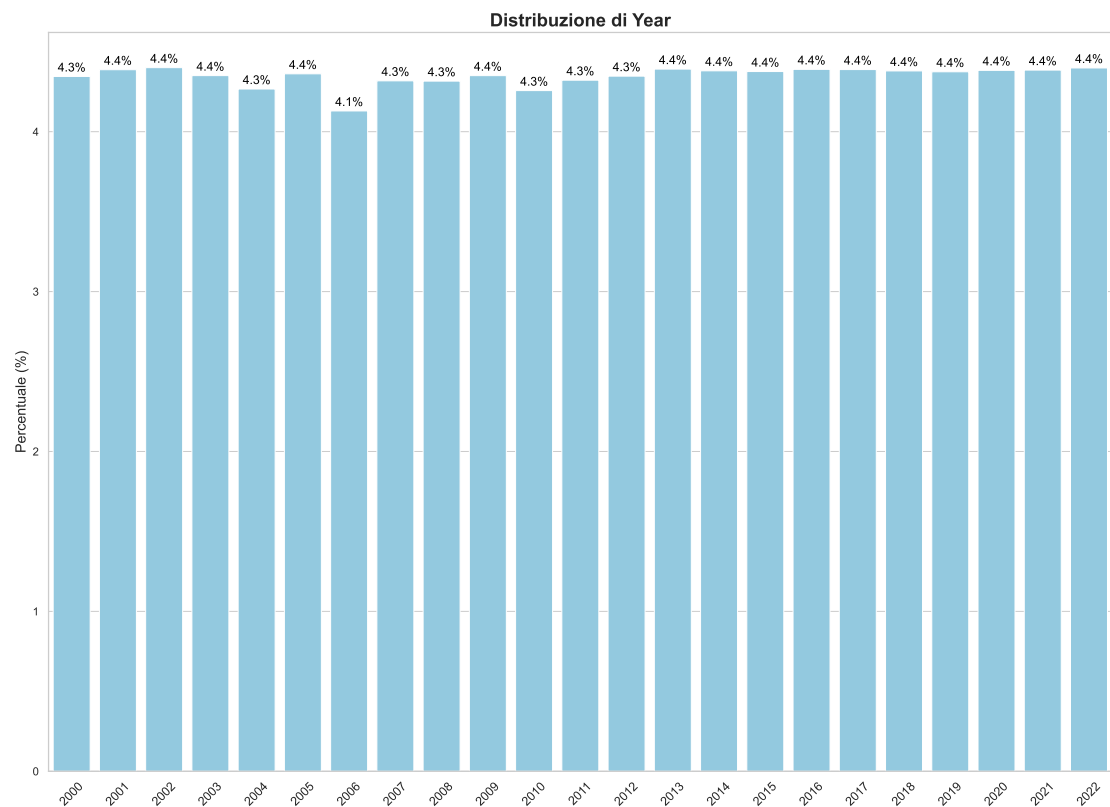



Figura 3.9: Grafico a barre della distribuzione della colonna Year nel dataset MOZ_Sat.

*Capitolo 4*

# *Schema e domande di ricerca*

Il problema è stato approcciato con la metodologia illustrata in Figura 4.1. A sinistra, in blu, troviamo le fonti, descritte dettagliatamente in precedenza, con le operazioni preliminari effettuate su di esse. Da queste ricaviamo dei risultati che, tramite diverse tecniche e processi di incrocio, rispondono alle domande di ricerca. Le domande di ricerca sono rappresentate in verde e hanno diversi collegamenti.

Inizialmente, è stato fornito l'elenco dei nomi dei dataset di interesse riguardanti disboscamento e malaria in Mozambico nel periodo dal 2000 al 2022. Questi corrispondono quindi a foreste, Land Cover e MAP dal sito Africa Knowledge Platform. Il primo obiettivo è stato quello di elaborare questi dataset e unirli in un unico file CSV per poi fornirlo per altre ricerche. Le tecniche di conversione e unione utilizzate e il risultato corrispondente a questa domanda sono già stati esposti in precedenza, nel Capitolo 3, ottenendo il dataset MOZ_Tree_Land_Malaria.

Il secondo obiettivo è stato quello di analizzare le fonti, introducendone anche di nuove, e rispondere alle domande di ricerca. In particolare, si è cercato di capire com'è il territorio in Mozambico, come cambia la vegetazione e quindi verificare se sono presenti aree dove la vegetazione diminuisce. Parallelamente, ci si è interrogati su come variano gli indici di malaria nel territorio e se si è verificato un aumento o una diminuzione dei casi nel tempo. Questo è stato ricercato sia a livello nazionale

Figura 4.1: Diagramma della metodologia di ricerca.

che in aree ristrette con raggio di alcuni chilometri, per comprendere l'influenza di fattori locali sull'incidenza della malaria che potrebbero non emergere da un'analisi a livello nazionale.

Unendo i risultati emersi, utilizzando nuove tecniche e introducendo nuove fonti, si è cercato di rispondere alle domande principali di questa ricerca: **c'è una relazione tra tipologia di terreno, in particolare vegetazionale, e infezioni malariche in Mozambico?** e, più specificamente, **c'è relazione tra diminuzione della vegetazione e aumento delle infezioni malariche?**

Inoltre, sono emerse nuove domande con lo scopo di fornire risultati per contribuire a ricerche condotte da dottorandi e futuri articoli scientifici. Per questo proposito è stato chiesto di considerare il periodo dal 2007 al 2022. Sono stati quindi forniti da dottorandi nuovi dataset riguardanti i villaggi e le miniere in Mozambico, già trattati in precedenza. L'interesse è quello di capire come varia il terreno attorno ai villaggi per verificare se influenza la trasmissione della malaria assieme ad altri fattori come gli interventi del governo. Inoltre, un'altra domanda riguarda la tipologia di terreno e le precipitazioni attorno alle miniere in Mozambico, per osservare se tali dati influenzano lo scatenarsi di conflitti da parte della popolazione nei pressi delle miniere. In aggiunta, è emersa un'altra domanda: se sono presenti aree all'interno della nazione dove la vegetazione diminuisce localmente ma non nelle aree circostanti. Questo per capire se esistono zone dove viene effettuato disboscamento per effetto diretto dell'uomo, come ad esempio per la costruzione delle miniere.

Le risposte a tutte queste domande di ricerca, le tecniche utilizzate e i risultati ottenuti verranno descritte nei capitoli successivi.

*Capitolo 5*

# *Tendenze vegetazionali e malariche, crescita e diminuzione*

La prima domanda si propone di analizzare le tendenze generali della vegetazione e della malaria considerate separatamente, per verificare se vi sia stata una diminuzione della vegetazione, indicando un cambiamento forestale significativo, e se viene osservata una tendenza all'aumento o alla diminuzione dei casi di malaria, sia a livello nazionale che locale.

## 5.1 Tecniche utilizzate

Inizialmente viene implementata una funzione per calcolare un indice di cambiamento della copertura forestale, chiamato *Forest Change Index* lungo tutta la sequenza temporale per ogni punto geografico. Questa verrà applicata al dataset MOZ_Tree_Land_Malaria, mancante delle colonne poco significative e dei valori nulli, che abbiamo visto precedentemente, e che contiene per ogni punto tutte le colonne Land Cover relative a ogni anno, che verranno utilizzate per il calcolo di questo primo indice. Questo indice sintetizza la variazione tra foresta e non-foresta su tutto l'intervallo di studio 2000-2022 utilizzando le colonne Land Cover

di copertura del suolo e classificandola in tre classi: 0 (stabile),1 (guadagno) e -1 (perdita). Viene utilizzato il linguaggio di programmazione Python e la libreria Pandas.

Prima di applicare la funzione, vengono divisi i possibili valori di copertura del suolo in due categorie foresta e non-foresta, seguendo le macrocategorie IPCC viste in Sezione 2.3.

La funzione prende come input una singola riga del DataFrame e itera sulla lista delle colonne Land Cover fino alla penultima. Per ogni coppia consecutiva di anni viene confrontato il valore di copertura del suolo dell'anno corrente con quello successivo. Se il valore corrente è classificato come foresta, ovvero è contenuto nella lista di valori che rappresentano la foresta, e il valore successivo appartiene alla lista non-foresta, viene decrementato l'indice di 1. Se invece avviene il contrario, da non-foresta a foresta, l'indice viene incrementato di 1. La funzione viene poi applicata a tutte le righe del DataFrame usando il metodo *apply* di Pandas con il parametro axis=1 (per lavorare riga per riga). Il risultato è memorizzato in una nuova colonna del DataFrame chiamata *Forest Change Index*.

Si passa ora dal dataset con coordinate univoche a un dataset che ha una riga per ogni punto geografico e anno. Questo viene fatto attraverso il metodo *melt*[1] di Pandas, che trasforma le colonne in righe, sistemando il contenuto della colonna *Year* e ordinando le righe in base alle tre colonne *Latitude*, *Longitude* e *Year*.

Si applicano ora due nuove funzioni per calcolare due nuovi indici legati alla copertura forestale. Viene effettuata un'analisi sia delle variazioni di stato della foresta nel tempo (colonna *Forest Change*) sia della proporzione cumulativa di anni in cui la copertura è classificata come foresta (colonna *Forest Mean*).

*Forest Change*: la funzione lavora su due righe consecutive, quella corrente e la precedente. Se la riga attuale è foresta e la precedente non lo è, sempre verificando se il valore è contenuto nelle rispettive liste, restituisce 1, indicando un guadagno di foresta. Se la riga attuale non è foresta e la precedente lo è, restituisce -1, indicando una perdita di foresta. Se non c'è un cambiamento (entrambe sono foresta o entrambe non lo sono) oppure non c'è una riga precedente (ad esempio per la prima riga del gruppo) restituisce 0. Il calcolo viene applicato a ogni gruppo di

[1] https://pandas.pydata.org/docs/reference/api/pandas.melt.html (visitato il 09/11/2024).

righe appartenenti allo stesso punto geografico, iterando per ogni gruppo scorrendo quindi gli anni di ogni riga in ordine cronologico per calcolare la variazione rispetto alla riga precedente.

*Forest Mean*: Questa colonna calcola la proporzione cumulativa di anni in cui un punto è stato classificato come foresta. Viene fatta un'iterazione per ogni gruppo di righe dello stesso punto e per ogni riga, se la copertura è foresta, un contatore viene incrementato. Parallelamente un'altro contatore totale viene incrementato per ogni riga. La proporzione cumulativa viene calcolata come $\frac{conteggioforesta}{conteggiototale}$ e memorizzata in una lista che viene aggiunta come colonna. Entrambe le funzioni hanno complessità computazionale in notazione O-grande[2] di $O(n)$.

Sempre sullo stesso DataFrame, si calcolano altre colonne che rappresentano il cambiamento dei diversi indici malarici nel tempo classificandoli in tre categorie 0,-1 e 1 sia a livello nazionale che locale. Questo viene fatto confrontando l'indice con la media della colonna e classificandolo nelle tre categorie: 0 se è vicino alla media (all'interno di una deviazione standard), -1 se è significativamente sotto la media (più di una deviazione standard al di sotto), 1 se è significativamente sopra la media (più di una deviazione standard al di sopra). Prendiamo in considerazione l'indice di *Incidence Rate*, prima a livello nazione e poi locale.

*Incidence Rate National Change*: Viene prima calcolata la media (*Dm*) e la deviazione standard (*Ds*) dell'indice nella sua intera colonna quindi in tutta la nazione e per tutti gli anni. Poi una funzione prende come input ogni riga e assegna la categoria restituendo 0 se il valore è compreso tra *Dm-Ds* e *Dm+Ds*, -1 se è inferiore a *Dm-Ds* e 1 se è superiore a *Dm+Ds*. Il calcolo viene applicato a tutte le righe del DataFrame. La complessità è $O(n)$.

*Incidence Rate Local Change*: Per analizzare i cambiamenti da un punto di vista locale, si confronta il valore di ogni punto con la media e la deviazione standard dei punti nell'area circostante. Per fare questo occorre, dato un punto, ricavare tutti i punti a esso vicini in un'area circolare di raggio definito in chilometri. A tal fine viene utilizzata una struttura dati chiamata *k-d tree*.

*K-d tree*: Il k-d tree, dove k è la dimensionalità dello spazio di ricerca, è stato introdotto da Jon Louis Bentley nel 1975 [3] ed è un albero di ricerca binario

[2] Notazione matematica utilizzata per descrivere il comportamento asintotico superiore di una funzione, caratterizzando la sua complessità in termini di crescita rispetto a un'altra funzione.

multidimensionale utilizzato per la ricerca associativa. È una struttura dati utile per l'archiviazione di informazioni da recuperare in modo efficiente tramite ricerche associative. Un albero binario è una struttura dati ad albero in cui ogni nodo può avere al massimo due figli: un figlio sinistro e un figlio destro. Parte da un nodo radice da cui si diramano i rami verso i figli fino ad arrivare alle foglie cioè nodi terminali senza figli. Nel k-d tree ogni nodo rappresenta un punto k-dimensionale attraverso un record di dati con k chiavi. A ogni livello l'albero opera su una dimensione specifica, iterando attraverso le k dimensioni. Ogni nodo ha un discriminatore associato, che è un intero compreso tra 0 e k-1. Il discriminatore determina quale chiave viene utilizzata per dividere i dati a quel livello dell'albero. Per ogni nodo $P$ in un k-d tree, se $j$ è il discriminatore di $P$, allora tutti i nodi nel sottoalbero sinistro di $P$ hanno una chiave j-esima minore della chiave j-esima di $P$, e tutti i nodi nel sottoalbero destro hanno una chiave j-esima maggiore della chiave j-esima di $P$. I discriminatori si alternano ciclicamente ad ogni livello dell'albero. La radice ha discriminatore 0, i suoi figli hanno discriminatore 1, e così via fino al livello k-esimo, dove il discriminatore è k-1. Al livello k+1-esimo il discriminatore torna a 0, e il ciclo si ripete. Questa struttura permette di eseguire operazioni di ricerca efficientemente tra cui la ricerca per regione, che mira a trovare record che intersecano una regione specificata nello spazio k-dimensionale, che useremo in questo caso, e la ricerca del vicino più prossimo che trova il record più vicino a un punto di *query* specificato, che utilizzeremo in seguito. La complessità in termini di tempo per l'inserimento di un nodo in un k-d tree è $O(\log n)$, dove $n$ è il numero di nodi, per la costruzione è $O(n \log n)$ utilizzando il metodo della mediana, per la ricerca del vicino più prossimo è $O(\log n)$ e per la ricerca per regione è $O(n^{1-1/k} + m)$ dove $m$ è il numero di nodi restituiti. La complessità in termini di spazio è $O(n)$.

Il k-d tree viene implementato attraverso la classe *KDTree*[3] della libreria *scipy.spatial* di Python che utilizza l'algoritmo descritto in [23] dove l'asse e il punto di divisione vengono scelti in base alla regola del *sliding midpoint*, che garantisce che le celle non diventino tutte lunghe e sottili. Inizialmente si estraggono dal DataFrame le coordinate geografiche univoche con il metodo *drop_duplicates* creando un nuovo DataFrame. Si crea quindi l'albero a due dimensioni passando alla clas-

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.html (visitato il 10/11/2024).

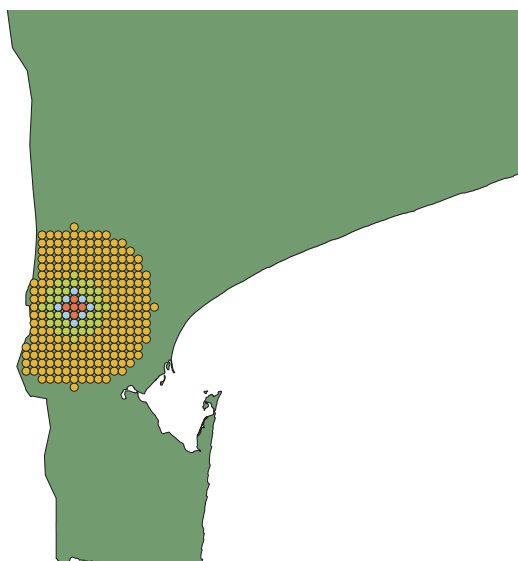



Figura 5.1: Confronto tra punti circostanti un dato punto con raggi di 5 km (rosso), 10 km (azzurro), 20 km (verde) e 50 km (arancione), ottenuti tramite la funzione *query_ball_point* del k-d tree.

se KDTree la latitudine e la longitudine di ogni punto. Ogni nodo rappresenterà quindi un punto con queste due coordinate, e la struttura suddivide ricorsivamente lo spazio bidimensionale per consentire ricerche rapide. Si definisce il raggio di ricerca in gradi geografici partendo da 0.0417 per una distanza equivalente a circa 5 km, che è la grandezza di ogni nostra cella. Per modificare e aumentare il raggio di ricerca basta modificare il valore di questa variabile moltiplicandola ad esempio per 2 per avere 10 km, 4 per 20 km e 10 per 50 km (Figura 5.1). Per ogni punto di coordinate univoche si trovano le coordinate dei punti vicini all'interno del raggio specificato utilizzando il metodo *query_ball_point*[4] della classe KDTree passando le due coordinate e il raggio. I risultati sono memorizzati in una lista di liste, dove ogni lista interna contiene gli indici dei punti vicini per ogni punto. Per ogni area definita dagli indici viene calcolata la media locale e la deviazione standard locale e vengono tutte memorizzate in due rispettive liste. Queste due liste vengono poi replicate per il numero totale di anni, in questo caso 23 poiché ogni coordinata

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.query_ball_point.html (visitato il 10/11/2024).

ha un valore per ogni anno. Per ogni punto del DataFrame il cambiamento locale viene calcolato confrontando il tasso di incidenza corrente con l'intervallo (media - deviazione standard, media + deviazione standard) e classificandolo nelle tre categorie 0, -1 e 1 come visto precedentemente. Prima di questo approccio era stato considerato un metodo *naive* di ricerca lineare che confrontava ogni punto con tutti gli altri, con complessità $O(n \cdot m)$ dove $n$ è il numero di punti del dataset e $m$ è il numero medio di vicini da trovare per ogni punto, che per dataset di grandi dimensioni come il nostro sarebbe stato molto inefficiente e l'esecuzione sarebbe risultata molto lenta. Con l'approccio del k-d tree, si ottiene, invece, una complessità che si può approssimare per molti dati di $O(n \log n)$.

Entrambe le colonne, nazionale e locale, vengono calcolate allo stesso modo per gli altri indici malarici del nostro studio ovvero *Parasite Rate* e *Incidence Count.*

Ora si analizza la tendenza a crescere o diminuire della vegetazione e della malaria in modo meno generale. Vengono presi in considerazione i cambiamenti di stato della foresta in base a più classi classificando i valori di Land Cover consultabili nella legenda in Sezione 2.3 in diverse categorie:

- Aree non forestali (0%): comprende i valori 11 (erba), 10-20 (terreni agricoli), 130 (praterie), 180 (vegetazione allagata), 190 (aree urbane), 120-121-122 (zone arbustive), 140 (licheni e muschi), 152-153 (vegetazione sparsa), e 200-201-202 (aree spoglie), 210 (corpi idrici);

- Altre aree con presenza di alberi: con i valori 12 (copertura arborea o arbustiva), 30 (mosaico di colture >50% con vegetazione naturale <50%), 40 (mosaico di vegetazione naturale >50% con colture <50%) e 110 (mosaico di copertura erbacea >50% con alberi e arbusti <50%);

- Foreste con copertura <15%: valori 150-151 (vegetazione sparsa con copertura arborea inferiore al 15%);

- Foreste con copertura >15%: con i valori 50 (latifoglie sempreverdi), 60 (latifoglie decidue), 70 (conifere sempreverdi) e 80 (conifere decidue) generali;

- Foreste con copertura tra 15-40%: valori 62 (latifoglie decidue aperte), 72 (conifere sempreverdi aperte) e 82 (conifere decidue aperte) tra il 15% e il 40% di copertura;

- Foreste con copertura >40%: include i valori 61 (latifoglie decidue chiuse), 71 (conifere sempreverdi chiuse) e 81 (conifere decidue chiuse) ovvero maggiori al 40% di copertura;
- Altre foreste: comprende i valori 90 (foreste miste), 100 (mosaico di alberi e arbusti >50% con copertura erbacea <50%), 160 (foreste allagate con acqua dolce o salmastra) e 170 (foreste allagate con acqua salata).

Ora il cambiamento della colonna Land Cover dalla riga precedente alla riga corrente viene classificato in base a queste categorie. La lista dei valori identificati come foresta è formata dalle categorie Foresta > 15%, Foresta tra 15% e 40%, Foresta > 40% e Altre foreste. Per ogni punto del DataFrame, raggruppandoli per latitudine e longitudine, e selezionando la riga corrente e precedente, il cambiamento di stato della copertura del suolo è identificato secondo la seguente logica:

- Cambiamento positivo (1): Si verifica quando c'è un passaggio da una categoria di non foresta o bassa copertura forestale (altre foreste o <15%) a una categoria di foresta, o un aumento di copertura forestale (ad esempio, da una foresta tra 15% e 40% a una foresta > 40%).
- Cambiamento negativo (-1): Si verifica quando avviene il contrario quindi c'è una diminuzione della copertura forestale, come in un passaggio da foresta a una foresta a bassa copertura o a non foresta.
- Nessun cambiamento (0): Quando la copertura del suolo rimane invariata o quando la riga precedente è nulla.

Si nota che i cambiamenti all'interno della stessa categoria, come ad esempio da foresta con copertura più bassa a foresta con copertura più alta, sono quasi nulli perché, come spiegato in Sezione 2.3, i cambiamenti vengono rilevati solo tra macrocategorie. Grazie a questa nuova colonna possiamo calcolare la tendenza complessiva della copertura forestale di ogni punto del Mozambico nel tempo, tornando quindi ad un DataFrame con una riga per punto geografico univoco. Lo scopo è quello di classificare il punto con classe 1 quando la copertura non diminuisce mai e aumenta almeno una volta, con classe -1 quando la copertura non aumenta mai e diminuisce almeno una volta, e con classe 0 in tutti gli altri casi. Si ottiene così una nuova colonna *Forest Trend* che rappresenta la tendenza della

copertura forestale. Entrambe le funzioni hanno complessità computazionale $O(n)$.

Per calcolare anche la tendenza della malaria negli anni per ogni punto univoco si utilizza la regressione lineare.

*Regressione lineare*: La regressione è un metodo di apprendimento supervisionato in cui l'obiettivo è imparare una funzione stimatore da dei dati di esempio. Questa funzione stimatore, chiamata anche regressore, è una mappatura da uno spazio di input a un valore reale ($f: \mathcal{X} \to \mathbb{R}$). Si usa la regressione, a differenza della classificazione dove abbiamo delle classi in numero finito, quando si vuole prevedere un valore numerico continuo. L'addestramento di un modello di regressione, chiamato *fit*, consiste nel trovare i valori ottimali per i parametri del modello che minimizzano l'errore sui dati di addestramento dati in "pasto" al modello. Questi parametri determinano la forma della funzione stimatore che il modello utilizzerà per fare previsioni. Il processo di addestramento può essere visto come la "sintonizzazione" del modello per adattarsi ai dati. Una volta che il modello è stato addestrato, può essere utilizzato per prevedere il valore della variabile dipendente per nuovi dati, questo processo è chiamato *predict* [9, 19]. La regressione può essere funzionale o simbolica, la regressione funzionale è un insieme di tecniche e algoritmi che consentono di estrarre una funzione matematica che descrive un fenomeno, mentre la regressione simbolica è dedicata all'inferenza di una teoria logica. La regressione funzionale, che è molto più popolare e comune, può essere semplice come una regressione lineare o complessa come una rete neurale. D'altra parte, gli approcci simbolici tipici includono alberi decisionali e foreste casuali, che vederemo in seguito, e regressori basati su regole [22]. La regressione lineare presuppone che il fenomeno sottostante possa essere approssimato con una linea retta (o un iperpiano, nel caso multivariato). Questo modello è il più semplice e indubbiamente il più popolare, e fa parte dei modelli lineari, cioè che possono essere rappresentati graficamente con linee o piani. Questi modelli sono interessanti per la loro semplicità e stabilità e sono meno inclini all'*overfitting* rispetto ad altri modelli più complessi, il che significa che sono meno propensi ad adattarsi troppo ai dati di addestramento memorizzando il loro rumore e di conseguenza a generalizzare male su nuovi dati. Tuttavia, possono anche condurre a *underfitting*, cioè potrebbero non essere in grado di catturare la complessità dei dati. In sintesi, quindi, viene tracciata una linea retta che meglio si adatta ai dati, permettendo

di fare previsioni su nuovi dati. Una retta ha equazione $y = mx + q$ dove $m$ è il coefficiente angolare e $q$ l'intercetta sull'asse delle ordinate. Il coefficiente $m$ rappresenta la pendenza della retta e quindi la direzione e l'inclinazione della retta, ed è questo che verrà utilizzato da noi per interpretare la tendenza della malaria.

La regressione lineare viene implementata attraverso la classe *LinearRegression*[5] della libreria *scikit-learn*[6] di Python, una libreria open-source per l'apprendimento automatico. Questa adotta la regressione lineare ai minimi quadrati ordinari, che minimizza la somma residua dei quadrati tra i target osservati nel dataset e quelli predetti dall'approssimazione lineare. Per ogni punto del DataFrame con coordinate/anno, raggruppandoli per latitudine e longitudine, ogni gruppo rappresenta un punto geografico con i valori dell'indice malarico per tutti gli anni dal 2000 al 2022. Si pone come $x$, variabile indipendente, gli anni del gruppo e come $y$, variabile dipendente, i valori dell'indice malarico. Viene quindi istanziato il modello di regressione inizializzando la classe *LinearRegression* passando il parametro *fit_intercept=True* in modo che venga calcolata anche l'intercetta e non assuma che la retta passi per l'origine. Viene quindi fatto il *fit* del modello passando come parametri $x$ e $y$ e si ottiene il coefficiente angolare della retta attraverso l'attributo $coef_[0]$ del modello. Questo viene inserito in una lista che verrà poi aggiunta come colonna. Il significato è quindi che se il valore $m$ è positivo, c'è un'incremento dell'indice malarico nel tempo per il punto specifico, se è negativo c'è una diminuzione nel tempo e se è prossimo a 0 non c'è nessun cambiamento significativo.

## 5.2 Risultati

### 5.2.1 Vegetazione

Per la prima colonna *Forest Change Index* si ottiene la quantità di punti per ogni classe in Tabella 5.1.

Quindi si ha l'1.7% di punti con perdita di foresta, il quale equivale a circa 15.000 km$^2$ della superficie del Mozambico considerando che ogni cella ha un'area

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html (visitato il 11/11/2024).

[6] https://scikit-learn.org/stable/index.html (visitato il 11/11/2024).

| Classe | Numero di punti |
|---|---|
| -1 | 625 |
| 0 | 36967 |
| 1 | 502 |

Tabella 5.1: Numero di punti per ogni classe di *Forest Change Index*.

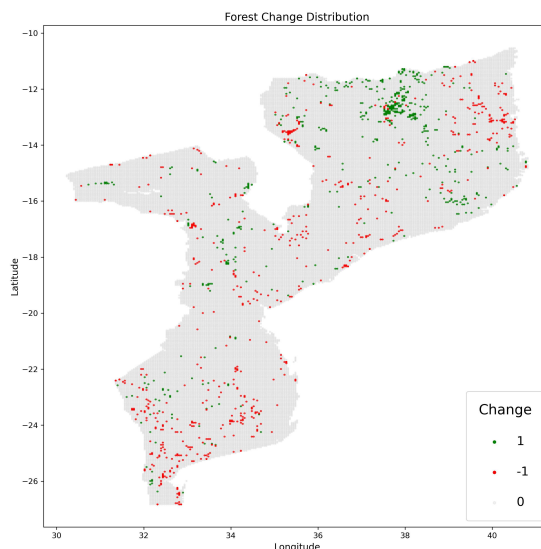



Figura 5.2: Grafico di dispersione dei punti di *Forest Change* sommati.

di 25 km$^2$ e l'1.3% di punti con guadagno di foresta.

Un esempio di un punto reale del dataset contenente le colonne *Forest Change* e *Forest Mean* discusse in precedenza è mostrato in Tabella 5.2. Vengono mappati questi punti e mostrati rispettivamente in Figura 5.2 e 5.3. Quest'ultima fa riferimento a tutti i punti dell'anno 2022 perché essendo la media cumulativa, viene visualizzata la storia completa di ogni punto considerando l'ultimo anno. La media della colonna *Forest Mean* è 0.6026 con deviazione standard 0.4855.

Si osservano inoltre i punti dell'indice *Forest Trend* in Figura 5.4 e la distribuzione in Tabella 5.3. Si nota un leggero aumento dei punti classificati come crescita o perdita di foresta rispetto alla colonna *Forest Change Index* che considerava solo la variazione macroscopica tra categorie IPCC.

| **Latitude** | **Longitude** | **Year** | **Land Cover** | **Forest Change** | **Forest Mean** |
|---|---|---|---|---|---|
| -26.8233 | 32.3125 | 2000 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2001 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2002 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2003 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2004 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2005 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2006 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2007 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2008 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2009 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2010 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2011 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2012 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2013 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2014 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2015 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2016 | 62.0 | 0 | 1.0 |
| -26.8233 | 32.3125 | 2017 | 180.0 | -1 | 0.9444 |
| -26.8233 | 32.3125 | 2018 | 180.0 | 0 | 0.8947 |
| -26.8233 | 32.3125 | 2019 | 180.0 | 0 | 0.8500 |
| -26.8233 | 32.3125 | 2020 | 180.0 | 0 | 0.8095 |
| -26.8233 | 32.3125 | 2021 | 180.0 | 0 | 0.7727 |
| -26.8233 | 32.3125 | 2022 | 180.0 | 0 | 0.7391 |

Tabella 5.2: Evoluzione temporale di un punto geografico dal 2000 al 2022, mostrando le colonne *Forest Change* e *Forest Mean*, che rappresentano rispettivamente le variazioni annuali della copertura forestale e la proporzione cumulativa di anni in cui il punto è stato classificato come foresta.

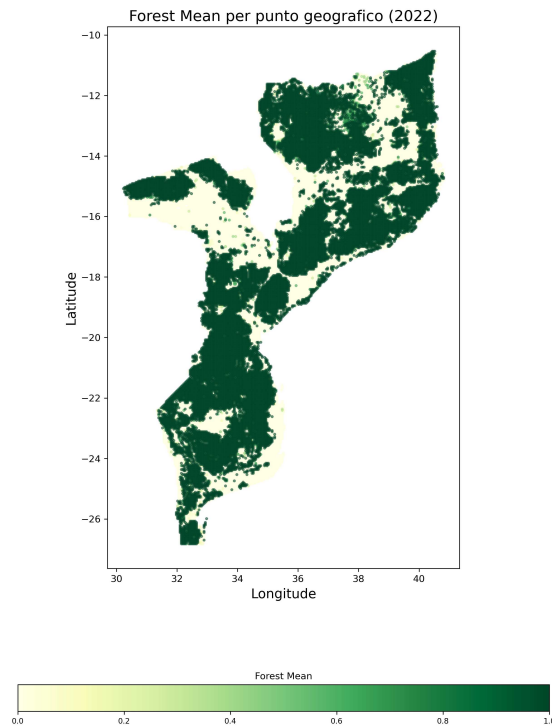



Figura 5.3: Grafico dei punti di *Forest Mean* nella mappa del Mozambico per l'anno 2022.

| **Classe** | **Numero di punti** |
|---|---|
| -1 | 630 |
| 0 | 36950 |
| 1 | 514 |

Tabella 5.3: Numero di punti per ogni classe di *Forest Trend*.

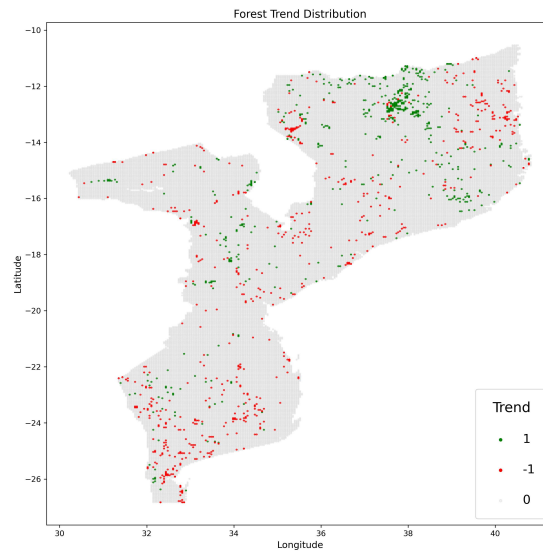



Figura 5.4: Grafico di dispersione dei punti di *Forest Trend*.

### 5.2.2 Malaria

Viene osservato un esempio di un punto evidenziando le colonne del cambiamento dell'incidenza della malaria a livello nazionale e locale in Tabella 5.4. La media dell'indice *Incidence Rate* a livello nazionale e per tutti gli anni usata per il confronto è 0.3954 con deviazione standard 0.1243. La media nazionale per l'indice *Parasite Rate* è 0.3720 con deviazione standard 0.1681 e per l'indice *Incidence Count* è 231.12 con deviazione standard 924.59. Viene ricavata la stessa tabella anche per gli altri indici malarici e per tutti i raggi locali descritti precedentemente.

I cambiamenti nel tempo per ogni anno degli indici calcolati per la metrica *Parasite Rate* sono mostrati in Figura 5.5 a livello nazionale e in Figura 5.6 a livello locale con raggio di 50 km. In verde sono rappresentati i punti che si trovano al di sotto della media nazionale e locale quindi classificati come -1, in rosso quelli al di sopra classificati come 1 e in giallo quelli vicini alla media classificati come 0.

In Tabella 5.5 si vede un esempio delle prime righe del dataset con le tendenze in cui le coordinate tornano univoche, da notare che l'ultimo punto è lo stesso esposto in Tabella 5.2 e 5.4, che ha subito una diminuzione forestale.

Si mappano i coefficienti angolari della regressione lineare per l'indice *Parasite Rate* in Figura 5.7, la media è -0.009191 con deviazione standard 0.005737. Mentre

| **Latitude** | **Longitude** | **Year** | **Incidence Rate** | **Inc Rate National Change** | **Inc Rate Local Change** |
|---|---|---|---|---|---|
| -26.8233 | 32.3125 | 2000 | 0.3108 | 0 | 0 |
| -26.8233 | 32.3125 | 2001 | 0.2646 | -1 | 0 |
| -26.8233 | 32.3125 | 2002 | 0.2489 | -1 | -1 |
| -26.8233 | 32.3125 | 2003 | 0.2238 | -1 | -1 |
| -26.8233 | 32.3125 | 2004 | 0.1736 | -1 | -1 |
| -26.8233 | 32.3125 | 2005 | 0.1300 | -1 | -1 |
| -26.8233 | 32.3125 | 2006 | 0.1062 | -1 | -1 |
| -26.8233 | 32.3125 | 2007 | 0.0969 | -1 | -1 |
| -26.8233 | 32.3125 | 2008 | 0.1112 | -1 | -1 |
| -26.8233 | 32.3125 | 2009 | 0.1671 | -1 | -1 |
| -26.8233 | 32.3125 | 2010 | 0.2463 | -1 | -1 |
| -26.8233 | 32.3125 | 2011 | 0.3028 | 0 | 0 |
| -26.8233 | 32.3125 | 2012 | 0.3187 | 0 | 1 |
| -26.8233 | 32.3125 | 2013 | 0.3067 | 0 | 0 |
| -26.8233 | 32.3125 | 2014 | 0.2607 | -1 | 0 |
| -26.8233 | 32.3125 | 2015 | 0.1966 | -1 | -1 |
| -26.8233 | 32.3125 | 2016 | 0.1556 | -1 | -1 |
| -26.8233 | 32.3125 | 2017 | 0.1410 | -1 | -1 |
| -26.8233 | 32.3125 | 2018 | 0.1311 | -1 | -1 |
| -26.8233 | 32.3125 | 2019 | 0.1389 | -1 | -1 |
| -26.8233 | 32.3125 | 2020 | 0.1819 | -1 | -1 |
| -26.8233 | 32.3125 | 2021 | 0.2115 | -1 | -1 |
| -26.8233 | 32.3125 | 2022 | 0.2189 | -1 | -1 |

Tabella 5.4: Evoluzione temporale di un punto geografico dal 2000 al 2022 del dataset MOZ_Land_Malaria_analysis, mostrando le colonne *Incidence Rate*, *Incidence Rate National Change* e *Incidence Rate Local Change*, la colonna locale fa riferimento a un raggio di 10 km. In questo caso la media locale è 0.2848 con deviazione standard 0.0265.

| **Latitude** | **Longitude** | **Forest Trend** | **Incidence Rate m** | **Parasite Rate m** | **Incidence Count m** |
|---|---|---|---|---|---|
| -26.823342 | 32.145833 | 0 | 0.001748 | 0.001859 | 0.045894 |
| -26.823342 | 32.187500 | 0 | 0.001081 | 0.001457 | -0.539462 |
| -26.823342 | 32.229167 | 0 | 0.000468 | 0.000882 | -0.937183 |
| -26.823342 | 32.270833 | 0 | -0.000577 | -0.000054 | -1.287588 |
| -26.823342 | 32.312500 | -1 | -0.001591 | -0.001096 | -1.233331 |

Tabella 5.5: Head del dataset MOZ_Land_Malaria_Trend_m.

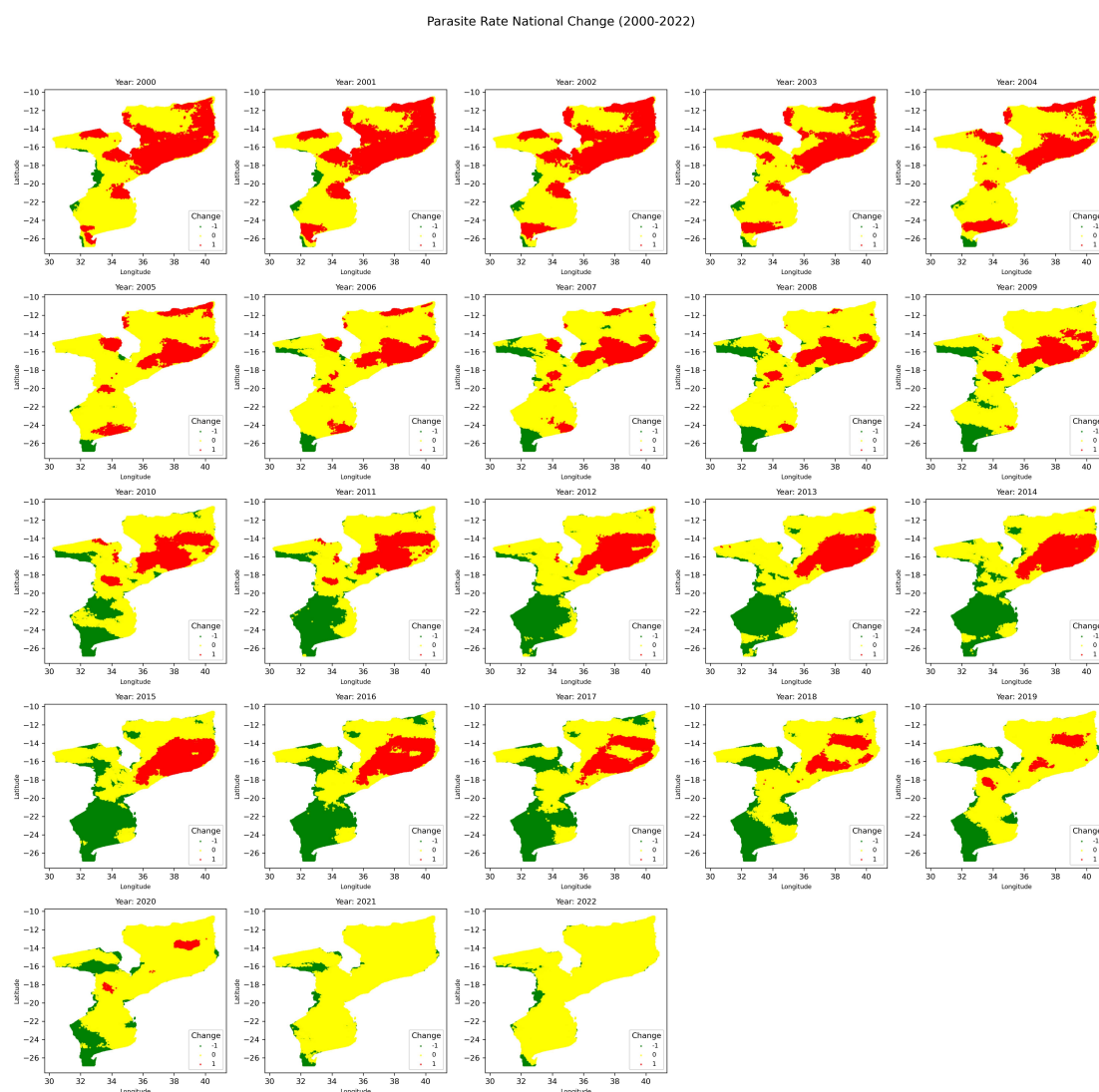



Figura 5.5: Grafici di dispersione della colonna *Parasite Rate National Change* per gli anni dal 2000 al 2022.

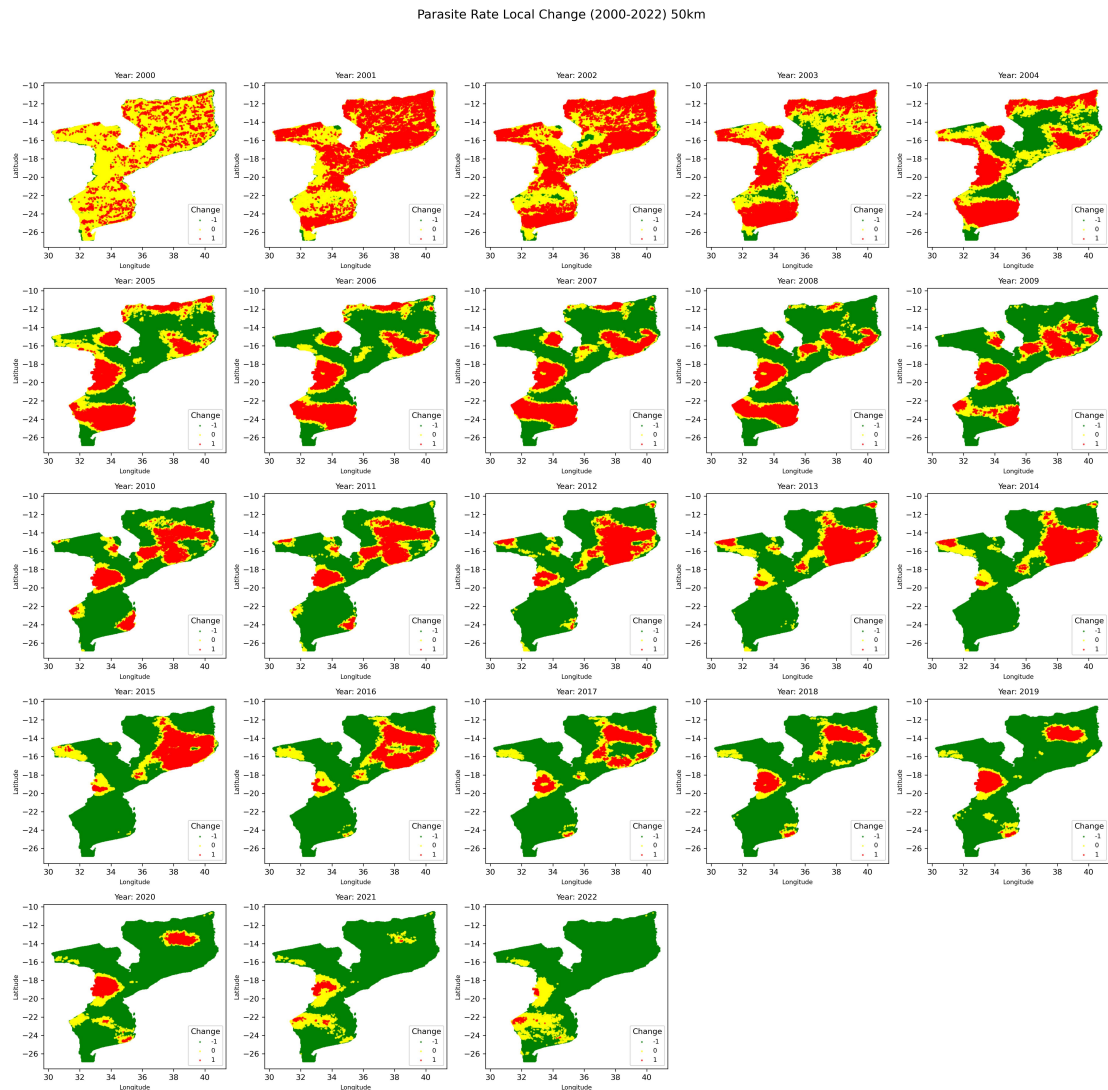



Figura 5.6: Grafici di dispersione della colonna *Parasite Rate Local Change* con raggio di 50 km per gli anni dal 2000 al 2022.

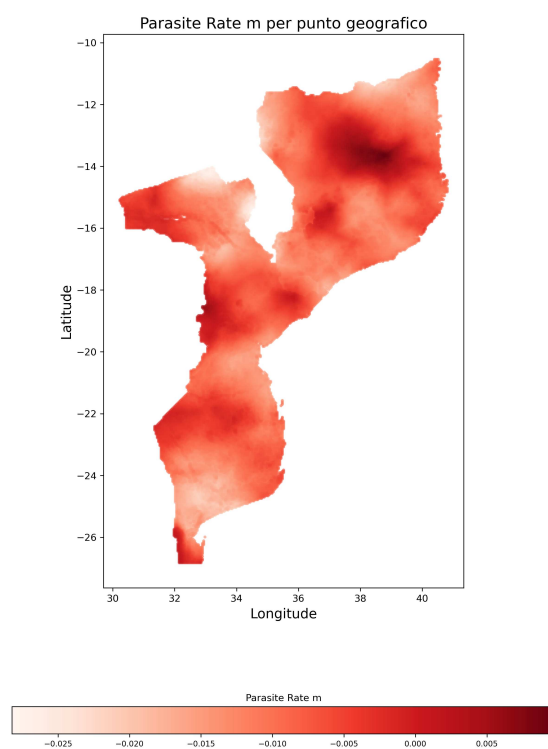



Figura 5.7: Mappa dei coefficienti angolari per l'indice *Parasite Rate*.

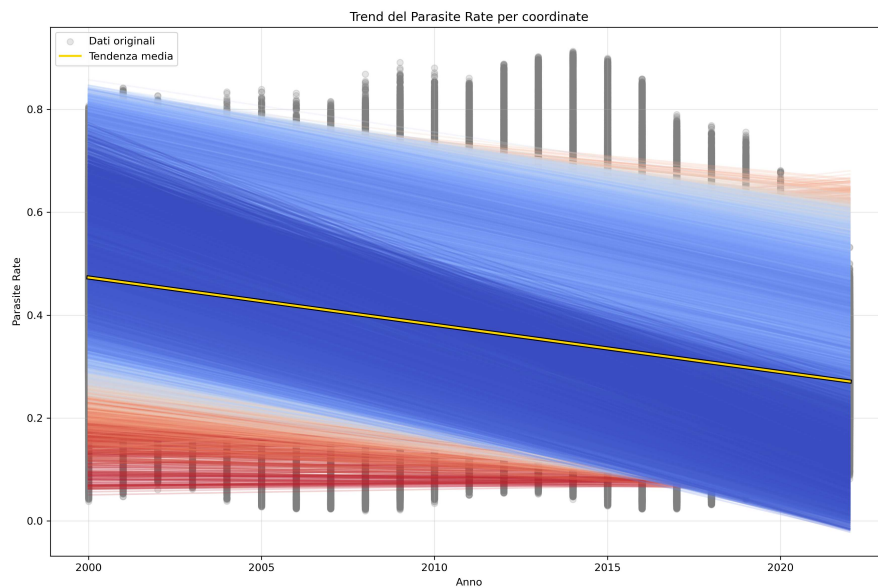



Figura 5.8: Grafico di tutte le rette di regressione con i punti originali per ogni anno in grigio, evidenziando la retta media.

un grafico che comprende tutte le rette di regressione generate per tutti i punti con anche la retta media è mostrato in Figura 5.8.

## 5.3 Risposta

Dall'analisi effettuata, emerge che nel periodo considerato (2000-2022) si è verificata una diminuzione più marcata della copertura forestale rispetto agli aumenti. Questo è evidenziato dai dati in Tabella 5.1 e 5.3. Parallelamente, l'analisi degli indici malarici mostra una tendenza generale alla diminuzione dei casi di malaria nel tempo. Come illustrato nelle Figure 5.5 e 5.6, si nota una evidente diminuzione dei punti classificati come -1 nel tempo, indicando una riduzione dell'incidenza della malaria. Inoltre, la regressione lineare effettuata ha mostrato un trend negativo come si evince dalla media negativa e dalla mappa in Figura 5.7. La Figura 5.8 conferma questa tendenza decrescente, mostrando che la retta di regressione media ha una pendenza negativa, evidenziando la diminuzione dell'indice sull'intero territorio analizzato. In conclusione, i risultati indicano che nel periodo studiato vi è stata sia una riduzione della copertura forestale sia una diminuzione dei casi di malaria in Mozambico.

*Capitolo 6*

# *Caratteristiche e variazioni territoriali attorno a villaggi e miniere*

In questo capitolo, come da richiesta, si cercano di unire i dati sul territorio con i dati forniti sui punti dei villaggi e sulle miniere. Inoltre, su richiesta, si aggiungono anche i dati relativi alle precipitazioni attorno alle miniere. Infine, analizzeremo anche se sono presenti perdite di foresta isolate sempre in relazione ad aspetti socioeconomici.

## 6.1 Tecniche utilizzate

Per ricavare il territorio attorno ai punti ricevuti sui villaggi e sulle miniere viene utilizzato il k-d tree per trovare il punto più vicino in un raggio determinato, aggiungendo alla determinata riga il nostro punto Land Cover che, rappresentando un'area di 5x5 km, coprirà il punto di interesse.

Per prima cosa si convertono le coordinate di entrambi i dataset in radianti per renderle compatibili tra di loro e assicurare la precisione dei calcoli sferici del

k-d tree. Viene quindi costruito l'albero a due dimensioni utilizzando la classe KDTree basato sulle coordinate in radianti del nostro DataFrame con tutti i punti Land Cover del Mozambico. Viene definito il raggio massimo in radianti, quindi ad esempio per 5 km sarà 5/6371, dove 6371 è il raggio della Terra in chilometri. Poi per ogni punto del DataFrame che contiene i punti dei villaggi, viene trovato il punto di Land Cover più vicino attraverso il metodo *query*[1] del KDTree. Vengono passati come parametri le coordinate del punto del villaggio, *k=1* per trovare un solo punto e *distance_upper_bound* per impostare il raggio massimo. Se il punto è stato trovato, viene unito alla riga del villaggio aggiungendo le coordinate del punto Land Cover più vicino, la distanza in chilometri e tutte le colonne Land Cover dal 2000 al 2022. La complessità della costruzione del KDTree è $O(n \log n)$ dove n è il numero di punti Land Cover, per la query per ogni villaggio è $O(\log n)$ e la complessità totale è quindi $O(n \log n) + O(m \log n)$ dove m è il numero di villaggi, ma nel nostro caso n è molto maggiore di m quindi possiamo considerare la complessità $O(n \log n)$. Questo approccio è molto efficiente nel caso di grandi dataset come il nostro dove la complessità *naive* sarebbe stata $O(n \cdot m)$ impiegandoci molto tempo per l'esecuzione.

Viene aggiunta anche una colonna per la prevalenza di tipologia di territorio negli anni, chiamata *Land Prevalence*, sia come numero che tradotta attraverso la legenda per essere meglio interpretata. Questo viene fatto con la funzione moda (*.mode()*) per ogni riga del DataFrame ottenuto precedentemente. Viene poi tradotta con una funzione che itera un dizionario.

Infine viene aggiunta una colonna *Forest Loss* che classifica il punto con 1,-1 e 0 con la stessa logica della colonna *Forest Trend* spiegata in Sezione 5.1 ma iterando gli anni del punto in una stessa riga invece che per ogni riga.

Si effettua poi lo stesso per i punti delle miniere, utilizzando la stessa tecnica con il k-d tree e le stesse colonne aggiuntive. Inoltre, vengono aggiunte le due colonne per le precipitazioni, *Average precipitation* e *Precipitation variability*, allo stesso modo, con il KDTree, ma utilizzando un raggio più ampio cioè di 22 km dato che queste rappresentano un'area di circa 22x22 km.

---

[1] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.KDTree.query.html` (visitato il 12/11/2024).

Per calcolare la perdita di foreste isolate viene utilizzato il dataset, con la colonna *Forest Trend* e coordinate univoche, creato precedentemente in Sezione 5.1 e il k-d tree. Il KDTree viene creato con le coordinate del dataset MOZ_Land_Malaria_Trend_m e viene definito il raggio variabile che va da 5 a 50 km. Per ogni riga si ottengono gli indici di tutti i punti vicini entro il raggio specificato con la funzione *query_ball_point* e vengono salvati in una lista di liste. Per ogni punto poi si analizza la sua lista di punti circostanti e, si verifica se il punto attuale ha avuto una perdita forestale (classe -1) e se tutti i punti circostanti non mostrano alcuna perdita (classe >= 0). In tal caso il punto viene classificato come perdita forestale isolata e viene classificato come -1, altrimenti viene classificato 0 indicando che non è una perdita isolata. Questo viene fatto per i raggi 5, 10, 20 e 50 km.

## 6.2 Risultati

La distribuzione dei punti dei villaggi e miniere nel Mozambico è mostrata in Figura 6.1.

(a) Punti miniere. (b) Punti sondaggi villaggi.

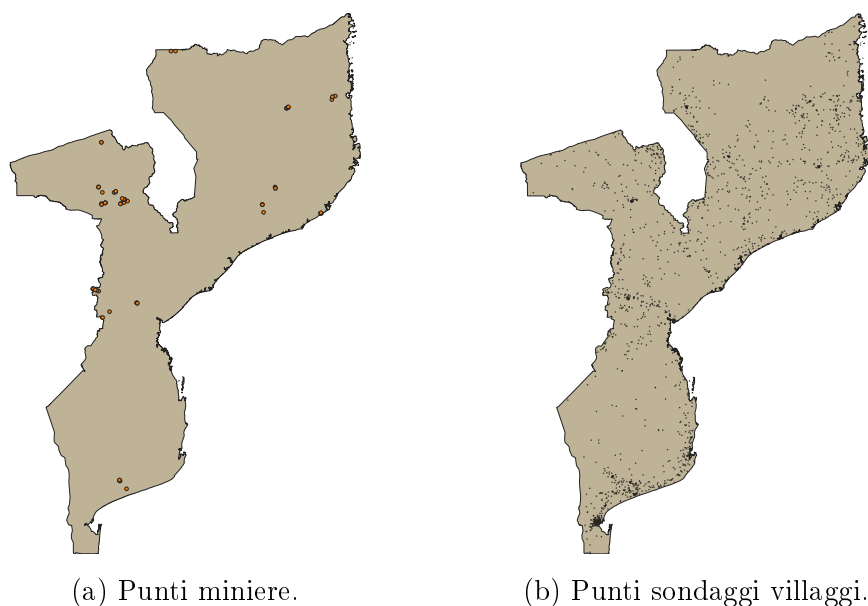



Figura 6.1: Mappe dei punti forniti delle miniere e dei villaggi in Mozambico.

| Vill Latitude | Vill Longitude | Closest Latitude | Closest Longitude | Distance | Land Cover 2007 | ... | Land Cover 2022 | Land Prevalence | Prevalenza | Forest Loss |
|---|---|---|---|---|---|---|---|---|---|---|
| -13.312592 | 35.246259 | -13.328518 | 35.229167 | 2.597711 | 11.0 | ... | 11.0 | Copertura erbacea | 0 | 0 |
| -13.282642 | 35.248191 | -13.286867 | 35.229167 | 2.167006 | 11.0 | ... | 11.0 | Copertura erbacea | 0 | 0 |
| -13.276335 | 35.273235 | -13.286867 | 35.270833 | 1.201102 | 11.0 | ... | 11.0 | Copertura erbacea | 0 | 0 |
| -13.301405 | 35.256039 | -13.286867 | 35.270833 | 2.306459 | 11.0 | ... | 11.0 | Copertura erbacea | 0 | 0 |
| -13.322513 | 35.248876 | -13.328518 | 35.229167 | 2.291068 | 11.0 | ... | 11.0 | Copertura erbacea | 0 | 0 |

Tabella 6.1: Head del dataset MOZ_Vill_Land_Analysis che analizza i punti dei villaggi.

| ... | LC 2022 | Land Prevalence | Prevalenza | Forest Loss | Closest Prec Lat | Closest Prec Long | Closest Prec Dist | Avg Prec | Prec Var |
|---|---|---|---|---|---|---|---|---|---|
| ... | 11.0 | 11.0 | Copertura erbacea | 0 | -16.524933 | 39.629480 | 3.550412 | 1148.950600 | 0.104109 |
| ... | 62.0 | 62.0 | Copertura arborea, latifoglie, decidua (15-40%) | 0 | -16.524933 | 37.827000 | 8.956602 | 1427.594800 | 0.102853 |
| ... | 11.0 | 11.0 | Copertura erbacea | 0 | -16.524933 | 39.629480 | 2.760222 | 1148.950600 | 0.104109 |
| ... | 11.0 | 11.0 | Copertura erbacea | 0 | -16.524933 | 39.629480 | 1.042252 | 1148.950600 | 0.104109 |
| ... | 11.0 | 11.0 | Copertura erbacea | 0 | -16.524933 | 39.629480 | 2.068227 | 1148.950600 | 0.104109 |

Tabella 6.2: Head del dataset MOZ_Mine_Land_Analysis che analizza i punti delle miniere e aggiunge i dati sulle precipitazioni, le prime colonne sono state omesse per brevità.

Le prime righe dei dataset ricavati per i villaggi e le miniere sono esposte in Tabella 6.1 e 6.2.

Si osserva la distribuzione delle prevalenze delle tipologie di copertura del suolo nei villaggi in Tabella 6.3. Il numero totale di villaggi è 2023 e si può notare che le Aree urbane costituiscono il 19.08% dei villaggi, mentre, considerando l'intera superficie del Mozambico, avevamo visto che costituivano soltanto lo 0.1% della superficie.

Per quanto riguarda la perdita di foresta nei villaggi, viene osservato in Tabella 6.4 il numero di punti per categoria. Sono 41 i villaggi su 2023 che hanno subito una perdita, lo 0.02%, e sono in numero maggiore rispetto a quelli che hanno avuto una crescita.

Anche per quanto riguarda le miniere, possiamo vedere in Tabella 6.5 la distribuzione delle tipologie di terreno.

Vengono inoltre confrontati i punti di perdita forestale isolata osservando come cambiano per ogni raggio di distanza in Tabella 6.6 e visivamente in Figura 6.2. Si nota che, come ci aspettavamo, il numero di punti diminuisce all'aumentare del raggio.

| Tipo di copertura | Prevalenza |
|---|---|
| Copertura erbacea | 402 |
| Aree urbane | 386 |
| Copertura arborea, latifoglie, decidua (15-40%) | 349 |
| Terreno coltivato, irrigato o post-inondazione | 211 |
| Arbusteto | 193 |
| Terreno coltivato, non irrigato | 129 |
| Copertura arborea, latifoglie, decidua (>15%) | 112 |
| Prateria | 58 |
| Copertura arbustiva o erbacea, inondata, acqua dolce/salata/salmastra | 44 |
| Mosaico di terreno coltivato (>50%) / vegetazione naturale (alberi, arbusti, copertura erbacea) (<50%) | 29 |
| Copertura arborea, inondata, acqua salata | 28 |
| Mosaico di alberi e arbusti (>50%) / copertura erbacea (<50%) | 22 |
| Corpi idrici | 22 |
| Mosaico di vegetazione naturale (alberi, arbusti, copertura erbacea) (>50%) / terreno coltivato (<50%) | 16 |
| Copertura arborea, latifoglie, sempreverde (>15%) | 15 |
| Aree spoglie | 5 |
| Mosaico di copertura erbacea (>50%) / alberi e arbusti (<50%) | 1 |
| Copertura arborea, latifoglie, decidua (>40%) | 1 |

Tabella 6.3: Prevalenza delle diverse tipologie di copertura nei villaggi.

| Forest Loss | Numero di punti |
|---|---|
| 0 | 1961 |
| -1 | 41 |
| 1 | 21 |

Tabella 6.4: Numero di punti per ogni classe di *Forest Loss* nei villaggi.

| Tipo di copertura | Prevalenza |
|---|---|
| Arbusteto | 22 |
| Copertura arborea, latifoglie, decidua (15-40%) | 16 |
| Copertura erbacea | 15 |
| Copertura arborea, latifoglie, decidua (>15%) | 13 |
| Terreno coltivato, non irrigato | 3 |
| Mosaico di copertura erbacea (>50%) / alberi e arbusti (<50%) | 1 |
| Terreno coltivato, irrigato o post-inondazione | 1 |
| Copertura arborea, latifoglie, sempreverde (>15%) | 1 |

Tabella 6.5: Prevalenza delle diverse tipologie di copertura nelle miniere.

| **Isolated Forest Loss** | **5 km** | **10 km** | **20 km** | **50 km** |
|---|---|---|---|---|
| 0 | 37714 | 37823 | 37969 | 38071 |
| -1 | 380 | 271 | 125 | 23 |

Tabella 6.6: Numero di punti per ogni classe di *Isolated Forest Loss* per ogni raggio di distanza.

(a) Raggio 5 km. (b) Raggio 10 km. (c) Raggio 20 km. (d) Raggio 50 km.

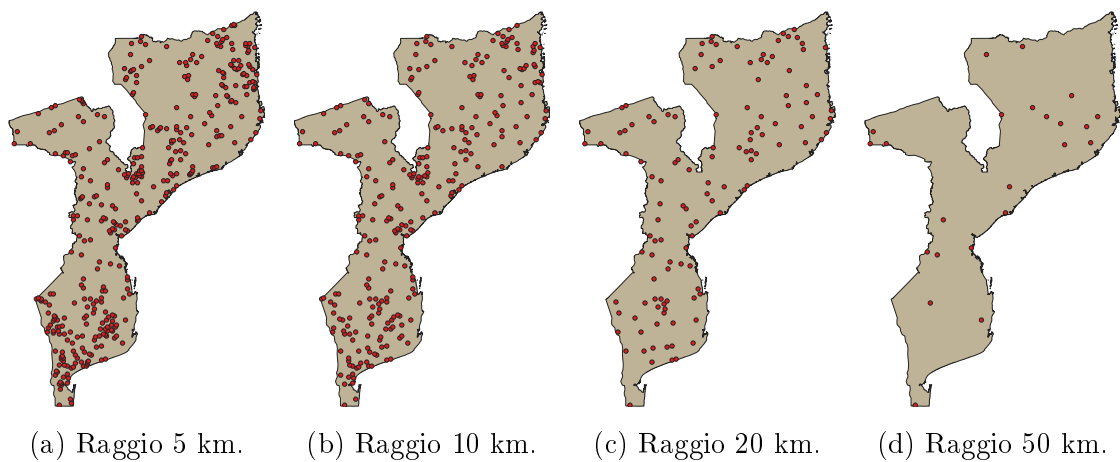



Figura 6.2: Mappe dei punti con perdita forestale isolata per ogni raggio di distanza.

## 6.3 Risposta

Le caratteristiche e le variazioni del territorio sono state espresse attraverso le tipologie di copertura del suolo, la presenza di perdita forestale e, per le miniere, anche attraverso le precipitazioni. Le classi di territorio trovate sono risultate coerenti con i villaggi e l'area di circa 5x5 km dei nostri punti Land Cover è risultata sufficiente per coprire i punti di interesse. La distanza tra i punti dei villaggi e miniere rispetto ai dati Land Cover e pluviometrici è risultata adeguata con punti sufficientemente vicini evitando distanze eccessive. Sia i dati di copertura territoriale che di precipitazione sono stati accettati e verranno utilizzati, assieme ad altri fattori, per studi sulla malaria e sui conflitti in Mozambico e ricerche in ambito socioeconomico. Possiamo quindi affermare che queste analisi condotte, assieme al dataset MOZ_Tree_Land_Malaria fornito precedentemente, hanno prodotto risultati significativi e utili per gli scopi prefissati e richiesti dai dottorandi.

*Capitolo 7*

# *Relazione tra tipologia di terreno e infezioni malariche*

In questo capitolo si analizza la relazione tra la copertura del suolo e le infezioni malariche, ovvero come la diffusione della malaria è influenzata dal tipo di terreno presente in un'area. Ad esempio se la presenza di foreste o di aree urbane possa favorire o sfavorire la diffusione della malattia e se ci possano essere differenze significative tra le varie tipologie di terreno. Questa relazione è stata studiata sia utilizzando la mappatura nazionale di Land Cover, sia analizzando i dati provenienti dai villaggi e, successivamente, integrando i dati satellitari.

## 7.1 Tecniche utilizzate

Inizialmente si classificano i punti del dataset avente serie temporale coordinata/anno in base alla colonna Land Cover in diverse categorie forestali, associando una percentuale di foresta seguendo le nuove categorie definite in Sezione 5.1. Per ogni riga del DataFrame viene estratto il valore Land Cover e, controllando all'interno di quale lista è contenuto, si assegna un valore di percentuale di foresta che

verrà inserito nella nuova colonna chiamata *Forest Category*. Le categorie della nuova colonna, in accordo con la classificazione precedente, sono:

- 0%
- Other Trees
- <15%
- >15%
- 15-40%
- >40%
- Other Forests

Poi si mettono in relazione queste categorie con gli indici malarici presenti nel DataFrame, raggruppando le righe per categoria forestale e visualizzando i valori dell'indice malarico corrispondente a ogni riga per ogni categoria attraverso un box plot (diagramma a scatola e baffi).

*Box plot*: Il box plot, introdotto da John Tukey nel 1970 [44], è uno strumento utile a visualizzare alcune delle statistiche rappresentative dei dati. Si ottiene sovrapponendo ad una linea orizzontale o verticale che va dal minore al maggiore dei dati, escludendo gli *outliers*, un rettangolo (il box) che va dal primo al terzo quartile dei dati, con una linea interna che lo divide al livello del secondo quartile. Un percentile è un dato che, in un insieme di dati numerici, per ogni valore $k$ indica che il $k$% dei dati è inferiore a quel valore e il *100-$k$%* è superiore. Un quartile è un percentile che divide i dati in quattro parti uguali, quindi il primo quartile è il 25-esimo percentile, il secondo quartile o mediana è il 50-esimo percentile e il terzo quartile è il 75-esimo percentile. Gli *outliers* sono i dati anomali identificati utilizzando un metodo che è funzione dello scarto interquantile (IQR) che è la differenza tra il terzo e il primo quartile e sono rappresentati come punti isolati al di sopra o sotto i baffi del box plot.

Il box plot viene implementato attraverso la libreria Seaborn[1] di Python con parametri $x$ la colonna *Forest Category* delle categorie forestali, $y$ l'indice malarico

[1] https://seaborn.pydata.org/ (visitato il 13/11/2024).

e un ordine personalizzato che va dalla percentuale minore alla maggiore.

Si passa poi all'analisi della relazione tra il territorio e la malaria attraverso i dati dei villaggi discussi precedentemente. Si uniscono al DataFrame MOZ_Vill_Land_Analysis, che contiene i punti dei villaggi e le informazioni sul territorio, i dati forniti sulla malaria e gli interventi del governo nei villaggi, provenienti da diverse fonti, facendo un *merge* di tipo inner su latitudine e longitudine. Le colonne aggiunte sono *year_test*, *tot_tested*, *test_positive_malaria*, *test_negative_malaria* e *gov_intervention* descritte in Sezione 2.6.

Vengono messe in relazione le categorie di prevalenza di copertura del territorio nei villaggi con il tasso di individui positivi alla malaria per ogni villaggio. Vengono escluse le colonne che hanno valore *tot_tested* uguale a 0, che quindi non hanno avuto sondaggi per la malaria, e viene calcolato il tasso di positività in una nuova colonna *Positive Rate* dividendo il numero di test positivi per il totale dei test effettuati. Osservando la tabella di distribuzione dei punti per ogni categoria di terreno si escludono le righe appartenenti a categorie con un basso numero di punti, quindi bassa granularità, mantenendo quindi aree significative e diverse per vegetazione. Le categorie mantenute corrispondono ai valori 11, 190, 62, 20, 120, 10, 60, 130, 180 e vengono messe in relazione attraverso il box plot con il tasso di positività alla malaria.

Per verificare se questa relazione è influenzata fortemente da altri fattori come la presenza di interventi governativi o l'anno di test, viene fatta un'analisi su di essi. Queste analisi vengono fatte sul DataFrame completo che esclude le righe senza test ma mantenendo tutte le categorie, e servono anche per analizzare la tendenza malarica in risposta alla domanda del Capitolo 5.

Vengono sostituite le celle vuote della colonna *gov_intervention* con la categoria di intervento *no intervention* per poterle considerare e analizzare assieme agli altri interventi. Dato che la colonna degli interventi esprime anche l'intervallo entro cui è stato effettuato l'intervento, ad esempio dal 2015 al 2018 o solo 2015, e abbiamo la colonna che indica l'anno in cui è stato effettuato il test malarico, si può notare che in alcuni casi l'intervento è stato effettuato dopo il test. Viene quindi estratto l'anno di inizio del trattamento manipolando le stringhe della colonna *gov_intervention* e salvato in una nuova colonna *start_year* e viene confrontato

con l'anno di test per sostituire la categoria dell'intervento in *treated after test* se l'intervento è stato effettuato dopo il test. Viene poi messa in relazione la presenza e il periodo di intervento con il tasso di positività alla malaria con il box plot.

Si dividono inoltre le righe in due categorie, quelle che hanno subito interventi e quelle senza interventi o dopo il test, con una nuova colonna con categorie *Yes* o *No* e anche queste due categorie vengono messe in relazione con il tasso di positività alla malaria. Viene messo inoltre in relazione l'anno di test con il tasso di positività per osservare se anch'esso incide.

Infine si divide il dataset in due DataFrame, uno con le righe che hanno avuto interventi e uno senza, e si mettono in relazione le tipologie di terreno significative selezionate precedentemente con il tasso di positività sia per il DataFrame con interventi che senza.

Per studiare questa relazione viene fatto uso anche dei dati derivati direttamente dalle immagini satellitari descritte ampiamente in precedenza. Si aggiunge al DataFrame contenente tutte le bande satellitari con punti coordinata/anno l'indice malarico *Parasite Rate*. Una volta estratte le colonne latitudine, longitudine e *Parasite Rate* dal DataFrame MOZ_Tree_Land_Malaria vengono modellate per avere le righe con punti coordinate/anno congruenti al DataFrame con le bande satellitari. Entrambi i punti, satellitari e malarici, rappresentano un'area di 5x5 km ma non è possibile unirli attraverso un merge diretto. Viene quindi utilizzato il k-d tree per trovare il punto malarico corrispondente al punto satellitare ma, a differenza delle precedenti iterazioni del k-d tree, le coordinate non sono univoche ma occorre gestire anche la corrispondenza con l'anno oltre che con le coordinate. Inizialmente le coordinate del DataFrame malarico vengono convertite in radianti e viene costruito il KDTree con esse. Per ogni riga del DataFrame con dati satellitari viene eseguita una funzione, attraverso *apply* e *lambda* per velocizzare ulteriormente il processo, con parametri le coordinate e l'anno del punto satellitare. Il punto ricevuto viene convertito in radianti e viene trovato il punto malaria più vicino usando *tree.query*, quindi ricerca nearest neighbour, restituendo l'indice e la distanza. Poi si verifica se l'anno del punto trovato coincide con l'anno desiderato e, nel caso contrario, viene cercato il punto più vicino con lo stesso anno e le stesse coordinate attraverso delle maschere booleane. Viene quindi restituita una serie contenente le coordinate del punto malaria più vicino, la distanza in chilometri

(usando il raggio terrestre di 6371 km) e il valore *Parasite Rate* associato. Successivamente si escludono dal DataFrame risultante le righe con distanza maggiore di 5 km con un filtro. L'ottimizzazione per questo problema attraverso il k-d tree, dove la ricerca ha complessità $O(\log n)$, è molto importante perché i due DataFrame, avendo le righe moltiplicate per tutti ventitre gli anni, hanno rispettivamente 777120 e 876162 righe che moltiplicandole risulterebbero oltre 680 miliardi.

Si utilizza ora questo dataset che viene chiamato MOZ_Sat_Mal per mettere in relazione le bande satellitari con l'indice malarico. Vengono escluse quindi le colonne non utili a questa analisi , ovvero le coordinate e la distanza del punto malaria, e viene costruita una matrice di correlazione su questo dataset.

*Matrice di correlazione*: Il coefficiente di correlazione è una misura della relazione tra due variabili. La correlazione tra due variabili può essere: positiva ovvero quando una variabile aumenta, anche l'altra tende ad aumentare; negativa dove quando una variabile aumenta, l'altra tende a diminuire; assente quando non esiste una relazione evidente tra le due variabili. Il coefficiente di correlazione più comune è quello di Pearson [28], calcolato come:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Dove $\mathrm{Cov}(X, Y)$ è la covarianza[2] tra le variabili $X$ e $Y$, e $\sigma_X$ e $\sigma_Y$ sono le deviazioni standard di $X$ e $Y$. Il risultato, $r$, varia da -1 a 1, con 1 che rappresenta una correlazione perfetta positiva, 0 che rappresenta assenza di correlazione e -1 che rappresenta una correlazione perfetta negativa. La matrice di correlazione di un dataset mostra i coefficienti di correlazione tra tutte le variabili a coppie.

Continuiamo l'analisi di questa relazione focalizzandoci solo sui punti relativi ai villaggi. Viene preso quindi il DataFrame con le coordinate dei villaggi e si moltiplica ogni riga per 23 aggiungendo la colonna *year* da 2000 a 2022 attraverso un prodotto cartesiano. Si abbina quindi ciascun villaggio con il punto satellitare più vicino, che contiene già l'indice malarico e le bande satellitari per uno specifico anno, utilizzando ancora il kd-tree. La distanza massima sarà sempre di 5 km e per ogni riga del DataFrame dei villaggi viene estratto l'anno e vengono selezionati i dati satellitari che corrispondono all'anno specifico saltando all'iterazione

[2] Misura statistica che indica la relazione di concordanza tra due variabili casuali, esprimendo quanto queste variano insieme, è calcolata come il valore atteso del prodotto delle loro deviazioni dalla media ($\mathrm{COV}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$).

successiva se non sono presenti dati satellitari disponibili per quell'anno. Con i dati satellitari filtrati viene costruito un KDTree utilizzando le coordinate e viene trovato il punto satellitare più vicino con la funzione *query* e la distanza massima di 5 km, se nessun punto viene trovato per quel villaggio e per quell'anno il punto viene ignorato. Si uniscono poi tutti i dati in un'unico DataFrame e anche su di questo, che si concentra sui villaggi, viene costruita una matrice di correlazione.

## 7.2 Risultati

La relazione tra malaria e categoria di foresta dell'area nell'estensione totale del Mozambico è esposta in Figura 7.1 con la relativa granularità delle categorie di foresta in Figura 7.2 per avere un'idea della distribuzione dei punti. Dopo lo 0% di foresta, le categorie con più punti sono le foreste con copertura tra il 15% e il 40% e le foreste con copertura più generale superiore al 15%.

Dopo aver filtrato le righe dei villaggi escludendo quelle con valore *tot_tested* uguale a 0, la prevalenza delle tipologie di copertura del suolo è cambiata ed è mostrata in Tabella 7.1. Questa è servita per decidere quali categorie più significative in base alla granularità mantenere per l'analisi successiva. Il numero di righe escluse rispetto al totale, 2023, è di 294. La relazione tra queste categorie di terreno mantenute e il tasso di positività alla malaria è mostrata in Figura 7.3. La media della colonna *Positive Rate* è 0.2741 con deviazione standard di 0.3060.

Dopo le trasformazioni effettuate sulle colonne degli interventi governativi, la distribuzione dei punti per categoria di intervento è mostrata in Tabella 7.2. Si osserva che tutte le righe con almeno un testato hanno ricevuto un intervento, dato che non è presente la categoria *no intervention*. Il box plot relativo è presentato in Figura 7.4.

La relazione tra la presenza o assenza di intervento governativo e il tasso di positività alla malaria è mostrata in Figura 7.5 e la distribuzione dei punti per categoria in Tabella 7.3. Le due categorie risultano simili per tasso di positività alla malaria.

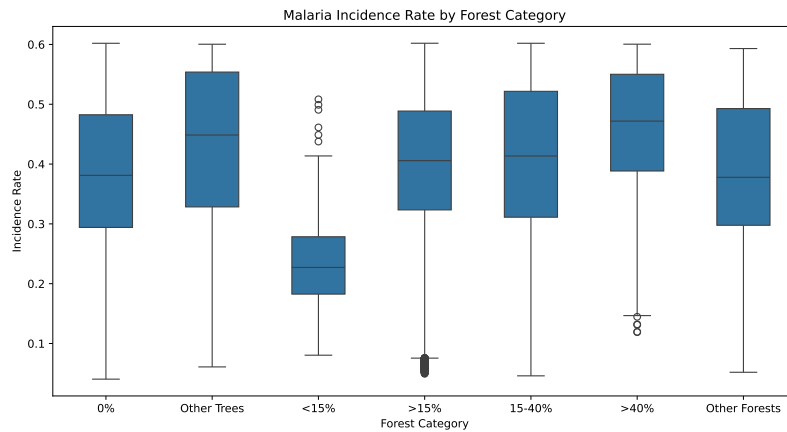
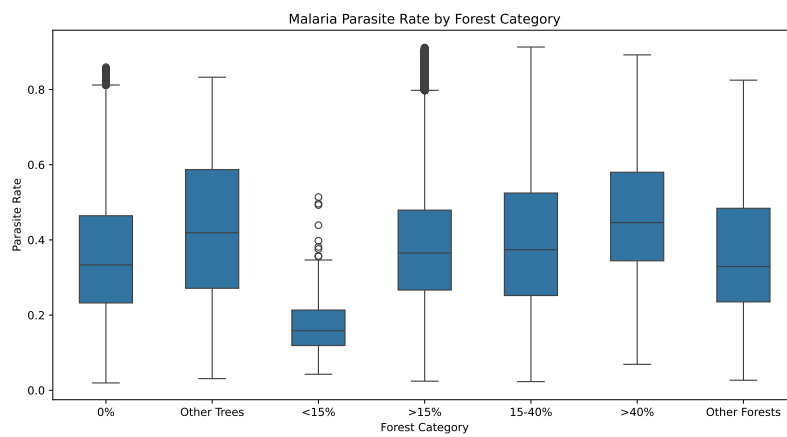
Viene esposta anche la relazione tra l'anno in cui è stato effettuato il sondaggio sulla malaria e il tasso di positività in Figura 7.6 e la distribuzione dei punti

(a) Incidence Rate.

(b) Parasite Rate.

Figura 7.1: Box plot della malaria per categoria di foresta.

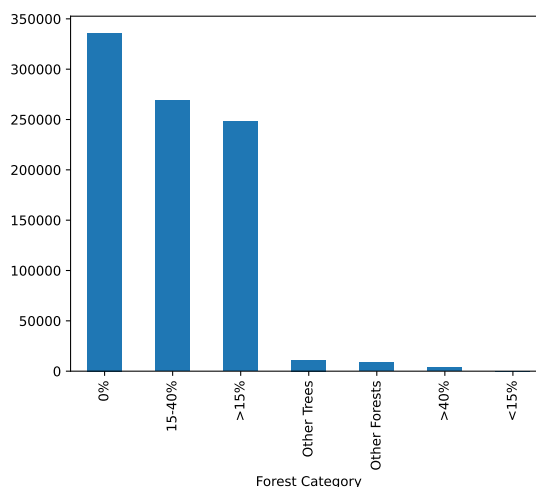



Figura 7.2: Distribuzione dei punti per categoria di foresta.

| Categoria | Tipo di copertura | Prevalenza |
|---|---|---|
| 11.0 | Copertura erbacea | 346 |
| 190.0 | Aree urbane | 319 |
| 62.0 | Copertura arborea, latifoglie, decidua (15-40%) | 307 |
| 20.0 | Terreno coltivato, irrigato o post-inondazione | 181 |
| 120.0 | Arbusteto | 166 |
| 10.0 | Terreno coltivato, non irrigato | 113 |
| 60.0 | Copertura arborea, latifoglie, decidua (>15%) | 93 |
| 130.0 | Prateria | 51 |
| 180.0 | Copertura arbustiva o erbacea, inondata, acqua dolce/salata/salmastra | 35 |
| 170.0 | Copertura arborea, inondata, acqua salata | 24 |
| 30.0 | Mosaico di terreno coltivato (>50%) / vegetazione naturale (alberi, arbusti, copertura erbacea) (<50%) | 21 |
| 210.0 | Corpi idrici | 20 |
| 100.0 | Mosaico di alberi e arbusti (>50%) / copertura erbacea (<50%) | 19 |
| 40.0 | Mosaico di vegetazione naturale (alberi, arbusti, copertura erbacea) (>50%) / terreno coltivato (<50%) | 15 |
| 50.0 | Copertura arborea, latifoglie, sempreverde (>15%) | 12 |
| 200.0 | Aree spoglie | 5 |
| 110.0 | Mosaico di copertura erbacea (>50%) / alberi e arbusti (<50%) | 1 |
| 61.0 | Copertura arborea, latifoglie, decidua (>40%) | 1 |

Tabella 7.1: Prevalenza delle diverse tipologie di copertura nei villaggi, escludendo quelli in cui non è stato effettuato alcun test.

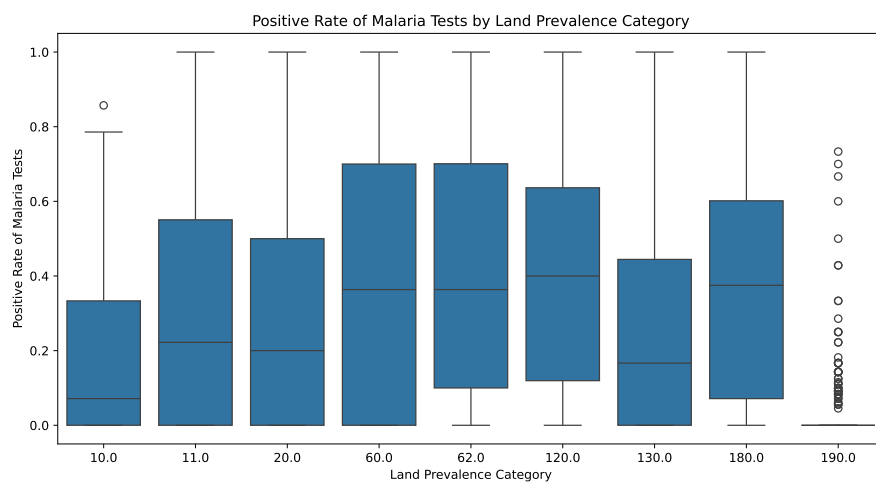



Figura 7.3: Box plot della malaria per categoria di copertura del suolo prevalente nei villaggi.

| **Intervento governativo** | **Numero di punti** |
|---|---|
| Treated after test | 519 |
| Treated from 2015 to 2023 | 513 |
| Treated from 2011 to 2023 | 336 |
| Treated only in 2015 and 2018 | 157 |
| Treated only in 2015 | 87 |
| Treated from 2018 to 2023 | 62 |
| Treated only in 2018 | 24 |
| Treated from 2011 to 2022 | 22 |
| Treated from 2023 | 9 |

Tabella 7.2: Distribuzione dei punti per categoria di intervento governativo.

| **Intervento** | **Numero di punti** |
|---|---|
| Yes | 1210 |
| No | 519 |

Tabella 7.3: Distribuzione dei punti in base alla presenza o assenza di intervento governativo.

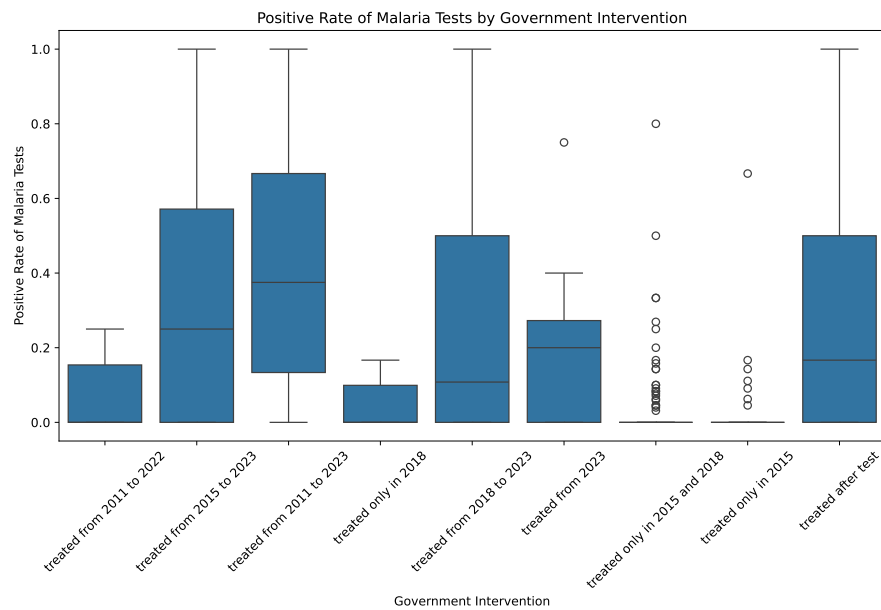



Figura 7.4: Box plot della malaria per categoria di intervento governativo.

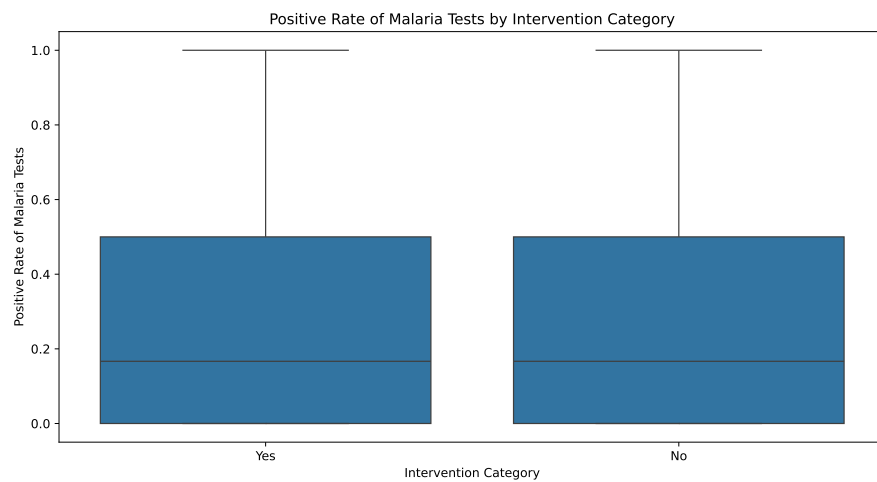



Figura 7.5: Box plot della malaria per presenza o assenza di intervento governativo.

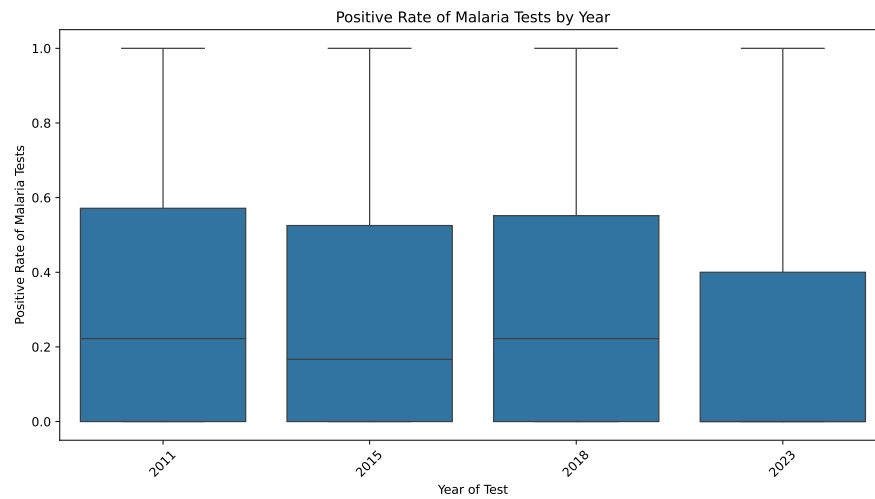



Figura 7.6: Box plot della malaria per anno di test.

| **Anno del test** | **Numero di punti** |
|---|---|
| 2011 | 603 |
| 2023 | 601 |
| 2015 | 304 |
| 2018 | 221 |

Tabella 7.4: Distribuzione dei punti per anno di test.

| Latitude | Longitude | Year | SR_B1 | SR_B2 | SR_B3 | SR_B4 | SR_B5 | SR_B7 | NDVI | EVI | Parasite_Rate |
|---|---|---|---|---|---|---|---|---|---|---|---|
| -26.837169 | 32.137229 | 2000 | 8201.5 | 9066.0 | 8963.5 | 15422.5 | 12679.0 | 9629.0 | 0.264823 | 2.257521 | 0.095427 |
| -26.837169 | 32.137229 | 2001 | 8629.0 | 9256.5 | 9411.0 | 14011.5 | 13261.5 | 10533.0 | 0.209266 | 2.077461 | 0.078630 |
| -26.837169 | 32.137229 | 2002 | 8652.0 | 9309.0 | 9563.0 | 13419.5 | 12734.0 | 10543.5 | 0.154688 | 1.537809 | 0.082713 |
| -26.837169 | 32.137229 | 2003 | 8218.0 | 9096.0 | 9260.0 | 14852.0 | 12876.0 | 10184.0 | 0.234529 | 1.587878 | 0.082074 |
| -26.837169 | 32.137229 | 2004 | 8191.0 | 8833.0 | 8673.0 | 14678.0 | 12097.0 | 9454.0 | 0.257917 | 2.437255 | 0.069532 |

Tabella 7.5: Head del dataset MOZ_Sat_Mal contenente dati satellitari e *Parasite Rate*.

per anno di test in Tabella 7.4. La positività sembra risultare stabile con una diminuzione nell'ultimo anno.

Infine, la relazione mostrata precedentemente in Figura 7.3 viene riproposta in Figura 7.7 divisa per presenza o assenza di intervento governativo. Si osserva una diminuzione generale dei casi in quasi tutte le categorie quando è presente un intervento governativo per contrastare la malaria.

I primi cinque record del dataset MOZ_Sat_Mal che unisce le bande delle immagini satellitari per ogni anno con l'indice malarico sono mostrati in Tabella 7.5 con la relativa matrice di correlazione in Figura 7.8.

Infine, la matrice di correlazione che unisce dati satellitari e indice malarico per tutti i punti corrispondenti ai villaggi è mostrata in Figura 7.9, ignorando per ora la colonna *cluster*.

## 7.3 Risposta

Dall'analisi condotta emerge che la prevalenza della malaria è più alta nelle aree con maggiore copertura forestale. Inizialmente nell'analisi della completa superficie del Mozambico basata sulla classificazione LC e sugli indici malarici del MAP (Figura 7.1) si osserva che la malaria è leggermente maggiore nei punti classificati come foreste soprattutto quando la copertura è maggiore del 15% e tra il 15% e il 40%, escludendo la categoria con copertura minore del 15% vista la bassa granularità. Questo risultato è confermato nell'analisi dei villaggi (Figura 7.3) dove si osserva che i casi di malaria sono significativamente maggiori dove c'è vegetazione, soprattutto forestale (categorie 60 e 62), rispetto alle aree urbane (categoria 190). Inoltre, nonostante si riscontri una diminuzione generale dei casi

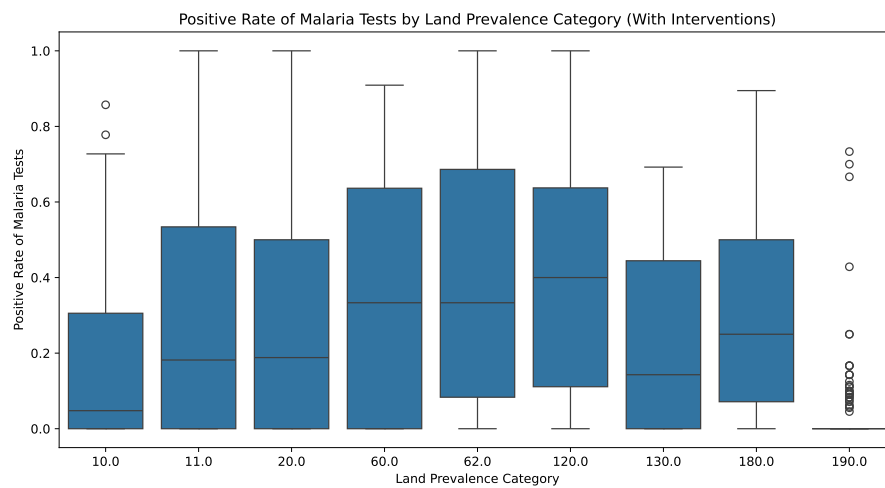
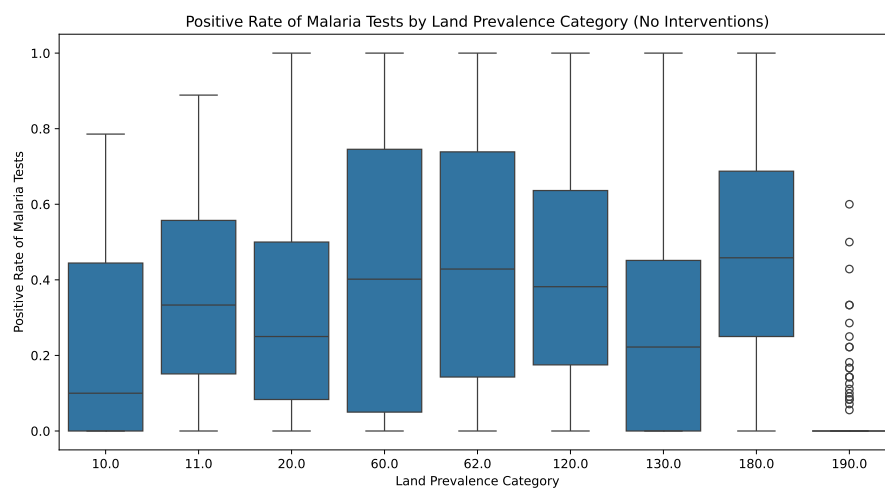
in quasi tutte le categorie a seguito di interventi mirati, persiste la differenza tra le aree urbane e quelle caratterizzate da vegetazione e foreste anche considerando gli interventi eseguiti dal governo (Figura 7.7). Viene trovato riscontro anche nell'analisi della correlazione tra i dati satellitari e l'indice malarico (Figura 7.8) dove emerge una correlazione moderata positiva tra il NDVI, indice vegetazionale, e il *Parasite Rate* di 0.23 generale e soprattutto nei villaggi (Figura 7.9) con un valore di 0.32, coerentemente con quanto già osservato nei risultati precedenti.

(a) Con intervento.

(b) Senza intervento.

Figura 7.7: Box plot della malaria per categoria di copertura del suolo prevalente nei villaggi con e senza intervento governativo.

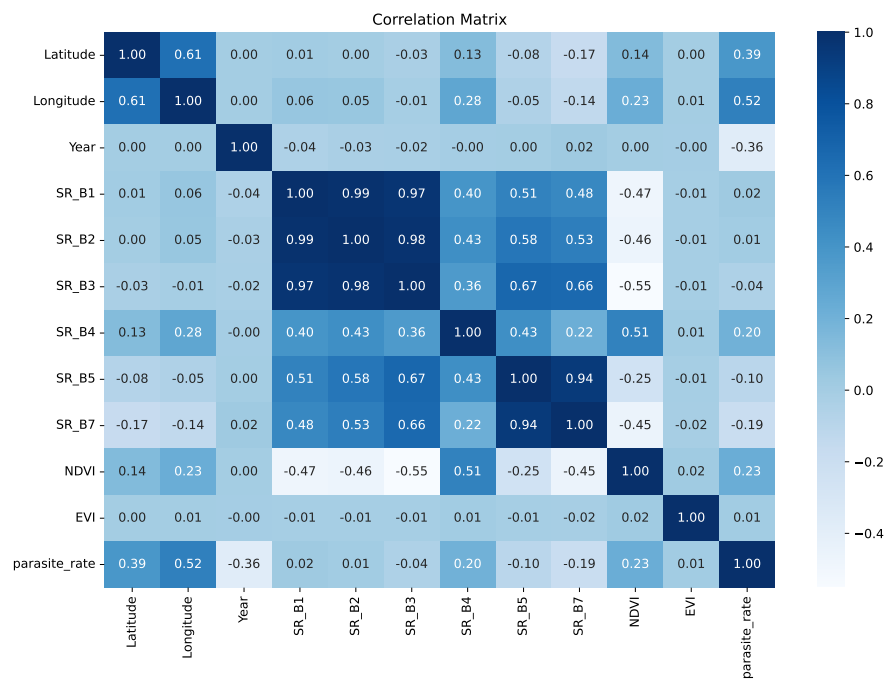



Figura 7.8: Matrice di correlazione del dataset MOZ_Sat_Mal.

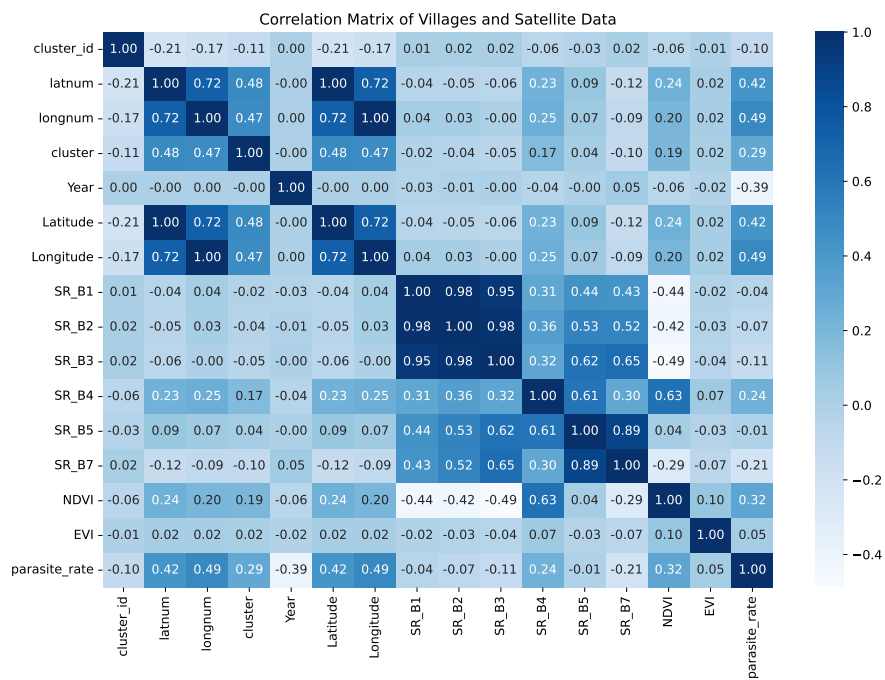



Figura 7.9: Matrice di correlazione del dataset che unisce i dati satellitari all'indice malarico per i villaggi.

*Capitolo 8*

# *Relazione tra diminuzione della vegetazione e aumento delle infezioni malariche*

In questo capitolo tratteremo più specificamente la relazione tra diminuzione della vegetazione e l'aumento delle infezioni malariche. Questa domanda ha suscitato l'interesse iniziale per questa ricerca e, pur essendo stata approfondita in altre zone del mondo come l'Amazzonia, non è ancora chiara la relazione tra deforestazione e malaria nel nostro territorio di studio, il Mozambico [2]. Anch'essa è stata analizzata combinando la mappa della copertura del suolo, gli indici malarici, i dati dei villaggi e le immagini satellitari.

## 8.1 Tecniche utilizzate

Per analizzare la relazione tra la diminuzione della vegetazione e l'aumento delle infezioni malariche, viene innanzitutto messa in relazione la colonna *Forest Trend* con classificazione 0, 1 e -1, descritta in Sezione 5.1, con gli indici malarici provenienti dal MAP raggruppando le righe per classe utilizzando il box plot.

Successivamente, si ripete l'analisi focalizzandoci sui punti relativi ai villaggi mettendo in relazione la colonna *Forest Loss*, ottenuta con la stessa logica della colonna *Forest Trend*, quindi utilizzando il cambiamento delle categorie LC e classificandola in 0, 1 e -1, con il tasso di positività alla malaria proveniente dai sondaggi nei villaggi. Anche questo facendo uso del box plot.

Passando alle immagini satellitari, vengono utilizzati i modelli di regressione *Decision Tree* e *Random Forests* e la metrica di performance Root Mean Square Error (RMSE) per valutare la relazione tra le bande satellitari e l'indice malarico.

*Decision Tree*: I *decision tree* sono modelli di apprendimento supervisionato utilizzati sia per la classificazione che per la regressione. La classificazione simbolica supervisionata è la strategia di *machine learning* (apprendimento automatico) per l'estrazione di una teoria esplicita (logica) che descrive un insieme di dati etichettati. Si contrappone alla classificazione funzionale supervisionata, che comprende ad esempio la regressione lineare vista precedentemente. Esistono molti metodi di classificazione simbolica, che possono essere distinti in modelli basati su alberi e basati su regole. I *decision tree*, alberi decisionali, sono modelli di classificazione ad albero e la loro introduzione risale a John Ross Quinlan in [33] e [32]. Un *decision tree* è composto da nodi e rami, ogni nodo interno rappresenta un test su una caratteristica, mentre ogni ramo rappresenta un risultato del test. Le foglie dell'albero rappresentano un'espressione logica, che è la congiunzione dei valori incontrati sul percorso dalla radice alla foglia. Ogni foglia ha un'etichetta di output: una classe per la classificazione o un valore numerico per la regressione. Il problema di estrarre l'albero decisionale ottimale, ovvero con un numero minimo di nodi, da un insieme di dati è NP-hard (tempo esponenziale) quindi l'apprendimento è realizzato con euristiche basate sull'entropia e sulla misura dell'impurità dei nodi. Noi utilizzeremo una versione ottimizzata dell'algoritmo Classification and Regression Trees (CART) [5], un'evoluzione del C4.5 di Quinlan, che supporta le variabili target numeriche, quindi la regressione, non formula regole decisionali esplicite e costruisce alberi binari utilizzando la *feature* e la soglia che producono il maggior *information gain* in ogni nodo. Nei decision tree regressori, la variabile target è continua e ogni foglia è etichettata con un valore numerico. L'obiettivo è costruire un albero che minimizzi la varianza dei valori target nelle foglie e

quindi la misura di impurità è sostituita dalla varianza. L'obiettivo è minimizzare la varianza su tutte le possibili suddivisioni di un dato genitore, scegliendo la suddivisione con la varianza minima. I vantaggi del *decision tree* sono la velocità nell'addestrare e testare, la facilità di interpretazione e assenza di bias ma possono essere soggetti a overfitting e avere un'elevata varianza. [26, 36, 29]

*Random Forests*: La *random forest* [4] è un metodo di *ensemble learning* (apprendimento d'insieme), cioè metodi che costruiscono più modelli predittivi a partire da versioni adattate dei dati di addestramento e combinano le loro previsioni. In particolare, la *random forest* è un metodo di *bagging* (bootstrap aggregating), cioè genera modelli diversi campionando parti casuali del dataset di origine per poi combinare le previsioni di tutti i modelli attraverso un voto di maggioranza per la classificazione o una media per la regressione. Nel nostro caso faremo uso della regressione e la *random forest* generalizza i *decision tree* per ottenere regressori basati su più alberi anziché su uno solo. Oltre alla casualità nel dividere il dataset per creare i *bootstrap samples*, la *random forest* introduce ulteriore casualità, ovvero quando si divide ogni nodo, durante la costruzione dell'albero, la miglior divisione può essere trovata anche attraverso un sottoinsieme casuale delle *feature* e non su tutte, per evitare che un singolo attributo dominante influenzi la divisione. Avendo queste due fonti di casualità, la *random forest* è meno soggetta all'overfitting e riduce la varianza rispetto al singolo albero, a volte a costo di aumentare leggermente il bias. Con il *bagging* circa il 33% dei punti rimane inutilizzato per ogni campionamento e, con questi punti chiamati *out-of-bag* (OOB), si può stimare l'errore e valutare il modello in modo integrato senza il bisogno di un set di validazione separato. Questi insiemi di alberi tendono a essere quindi più performanti dei singoli alberi e, sebbene siano considerati al limite tra l'apprendimento simbolico e quello funzionale, la loro natura simbolica è ancora evidente perché come i singoli alberi, possono essere analizzati e discussi ed è possibile estrarre le regole e quindi interpretarli in maniera chiara. [24, 4, 29]

*RMSE*: La radice dell'errore quadratico medio (RMSE) è una misura della differenza tra i valori predetti da un modello e i valori osservati. Misura la deviazione standard dei residui ed è sempre non negativa, quindi più è vicina a 0, migliore è il modello, al contrario delle metriche per la classificazione. L'RMSE è calcolata come la radice quadrata del Mean Squared Error (MSE) che calcola la varianza dei residui, e per questo, come la varianza e la deviazione standard, ha la stessa

unità di misura della variabile target. La formula è quindi:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Ovvero la radice quadrata della media delle differenze quadratiche tra i valori osservati $y_i$ e i valori predetti $\hat{y}_i$ su tutti gli $n$ esempi nel set di test.

Viene utilizzata, per l'implementazione, la libreria *scikit-learn* in Python che per i *decision tree*, come detto, utilizza una versione ottimizzata dell'algoritmo CART e per le *random forests*, a differenza dello studio originale, combina i classificatori calcolando la media delle loro previsioni probabilistiche, anziché lasciare che ogni classificatore voti per una singola classe.

Per dividere il dataset in training e test set, viene utilizzata la tecnica della *10-fold cross validation*, che è una *K-fold cross-validation* con k uguale a 10. La *cross-validation* ripete il calcolo dell'addestramento e del test su diversi sottoinsiemi (o split o fold) scelti casualmente dal dataset originale. In particolare, la *10-fold cross-validation* divide in modo casuale i dati in 10 parti (fold) di uguale dimensione, e usa 9 parti insieme per l'addestramento e 1 parte per il test. Esegue quindi questa operazione 10 volte, utilizzando ogni volta una parte diversa come test set e le restanti come training set e valuta ogni volta le prestazioni. Alla fine, calcola la media delle prestazioni del set di test, nel nostro caso l'RMSE. Altri metodi di divisione del dataset per la valutazione del modello possono essere il *full-training* che può portare a un totale overfitting e il classico *train-test split* che divide il dataset in parti fisse.

Si selezionano quindi le *feature* creando una lista, con tutte le bande satellitari, da *SR_B1* a *EVI* e vengono selezionate le colonne del DataFrame MOZ_Sat_Mal relative a queste *feature* come variabili indipendenti (X) e la colonna *Parasite Rate* del DataFrame come variabile dipendente (y) per l'addestramento del modello. Viene creata la metrica di valutazione con la funzione *make_scorer* passando come parametri l'RMSE e il parametro *greater_is_better* uguale a False. Si istanzia il modello di regressione *decision tree* attraverso la classe *DecisionTreeRegressor*[1] e viene calcolato l'RMSE medio attraverso la funzione *cross_val_score* attraverso la *10-fold cross validation*. Viene fatto lo stesso per il modello di regressione

[1] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html (visitato il 16/11/2024).

*random forests* con la classe *RandomForestRegressor*[2] passando come parametro anche *n_jobs* uguale a -1 per sfruttare tutti i processori disponibili e avere un calcolo più veloce. Vengono utilizzati 100 alberi e tutte le *feature* per il modello. Inoltre, viene fatto uso del *bootstrap* ma non del *OOB score*.

Si fa lo stesso anche per il DataFrame che ha sempre le stesse colonne ma riguarda solo i punti dei villaggi, prima sul totale dei punti e poi raggruppandoli in *cluster* attraverso l'algoritmo di apprendimento non supervisionato *K-Means*.

*K-means*: L'apprendimento automatico si divide in due categorie principali: apprendimento supervisionato e apprendimento non supervisionato. Mentre l'apprendimento supervisionato, che abbiamo visto fino ad ora, si basa su dati etichettati per addestrare modelli predittivi, l'apprendimento non supervisionato si concentra sull'individuazione di schemi e strutture in dati non etichettati. Ci sono due tipi di apprendimento non supervisionato, quello predittivo e quello descrittivo. Noi ci concentreremo sul secondo, che mira a descrivere i dati in modo nuovo e scoprire informazioni nascoste, mentre il primo, predittivo, può essere applicato a nuovi dati. Un esempio chiave di apprendimento non supervisionato è il *clustering*, che mira a suddividere i dati in gruppi (cluster) omogenei in base alla loro similarità. La valutazione degli algoritmi di apprendimento descrittivo è meno immediata rispetto alla valutazione di quelli predittivi non essendoci una variabile target, per questo nel clustering si valuta una particolare partizione dei dati calcolando la distanza media dal centro del cluster. Esistono diverse tecniche di clustering, tra cui il clustering esclusivo dove ogni elemento può essere assegnato a un solo gruppo, come nel caso del *K-means*. Il problema del *K-means* divide un insieme di $n$ campioni in $k$ cluster disgiunti $c_i$ con $i = 1, \ldots, k$, ciascuno descritto dalla media $\mu_i$ dei campioni nel cluster. Queste medie sono comunemente chiamate centroidi del cluster. L'obiettivo è minimizzare la somma dei quadrati delle distanze tra i campioni e il centro del cluster cioè risolvere il seguente problema di minimizzazione:

$$\arg\min_{c} \sum_{i=1}^{k} \sum_{x \in c_i} \|x - \mu_i\|^2$$

[2] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (visitato il 16/11/2024).

dove $c_i$ è l'insieme di punti appartenenti al cluster $i$ e $\mu_i$ è il centro del cluster $c_i$

$$\mu_i = \frac{1}{|c_i|} \sum_{x \in c_i} x$$

La funzione obiettivo del clustering K-means utilizza il quadrato della distanza euclidea $\|x - \mu_j\|^2$, spesso chiamata inerzia o somma dei quadrati intra-cluster e descrive la compattezza degli elementi dentro al cluster. La questione è che questo problema è NP-hard, quindi impraticabile, e per questo si utilizzano euristiche come l'algoritmo di Lloyd [21] che tramite una procedura iterativa spera di trovare il minimo globale, ma potrebbe bloccarsi in una soluzione diversa. Questa procedura assegna iterativamente i punti al cluster più vicino e aggiorna i centri dei cluster, convergendo verso una soluzione che minimizza la varianza intra-cluster. Questo ha quindi il difetto di convergere a minimi locali, cioè soluzioni migliori localmente ma non globalmente, soprattutto a causa della sua sensibilità alla scelta iniziale dei centroidi. [9, 46, 27, 29]

Il *K-means clustering* viene implementato con la classe *KMeans*[3] di *scikit-learn* utilizzando l'algoritmo default di Lloyd e, dato che l'algoritmo richiede di specificare il numero di cluster, si sceglie di utilizzare 20 cluster per raggruppare i villaggi. Viene istanziata quindi la classe *KMeans* e si usa il metodo *fit_predict* passando le coordinate di tutti i villaggi per creare una nuova colonna cluster che assegna a ciascun villaggio il gruppo numerico a cui appartiene. Viene calcolato quindi l'RMSE con il *decision tree* e il *random forests* con *10-fold cross validation* per ogni *cluster* semplicemente ciclando su tutti i *cluster* e ripetendo la metodologia descritta precedentemente.

I risultati ottenuti, però, potrebbero essere ingannevoli, dato che, se si fa una regressione tra quantità che non variano molto, quindi nel nostro caso se il territorio non varia, si ottengono delle metriche di performance molto buone perché la varianza è molto bassa e non c'è molto da predire. Per verificare se questo è il caso, vengono cercate delle aree in cui i valori di vegetazione cambiano molto e si restringe l'analisi a queste aree. Per fare ciò viene utilizzata la banda *NDVI* che va da -1 a 1 ed è un indice di vegetazione. Si calcola quindi la deviazione standard dell'*NDVI* per ogni punto del DataFrame completo lungo tutti i suoi

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html (visitato il 16/11/2024).

anni e si salva in nuova nuova colonna *NDVI_std* in un nuovo DataFrame con coordinate univoche. Si selezionano quindi solo i punti con deviazione standard maggiore di 0.10, che rappresentano le aree con maggiore variazione di vegetazione lungo i propri anni. Si unisce poi il filtro con il DataFrame originale con un *merge* per ottenere solo i punti selezionati. Confrontando i punti filtrati con le immagini satellitari durante gli anni, si può verificare che coprono aree dove c'è stato un visibile cambiamento di vegetazione e, al contrario, si osserva un'assenza di punti nelle zone maggiormente urbane o preservate, come ad esempio i parchi naturali. Viene calcolata quindi la matrice di correlazione e l'RMSE per tutti i punti filtrati dell'intera area del Mozambico. Successivamente, si ripete l'analisi per i villaggi. Viene preso il dataset con le coordinate dei villaggi e vengono aggiunte le colonne del nuovo dataset che contiene i punti filtrati per ogni anno, usando il k-d tree. Impostando il raggio massimo a 5 km si ottengono solo i punti dei villaggi che subiscono una variazione di vegetazione. Vengono calcolate le metriche RMSE su tutti questi villaggi filtrati e vengono raggruppati ulteriormente in cluster, ma, essendo questa volta in numero minore, si sceglie di utilizzare 5 cluster. Infine viene calcolato l'RMSE anche per questi ultimi cluster filtrati.

## 8.2 Risultati

La relazione tra i coefficienti angolari degli indici malarici e la colonna *Forest Trend* generale per tutta l'area del Mozambico è presentata in Figura 8.1. La distribuzione dei punti per categoria della colonna *Forest Trend* è stata mostrata in Tabella 5.3.

In Figura 8.2 è esposta la relazione tra la colonna *Forest Loss* nei punti corrispondenti ai villaggi e il tasso di positività alla malaria ottenuto dividendo il numero di casi positivi per il numero di test effettuati nei sondaggi. La distribuzione delle categorie per la colonna *Forest Loss* era stata presentata in Tabella 6.4.

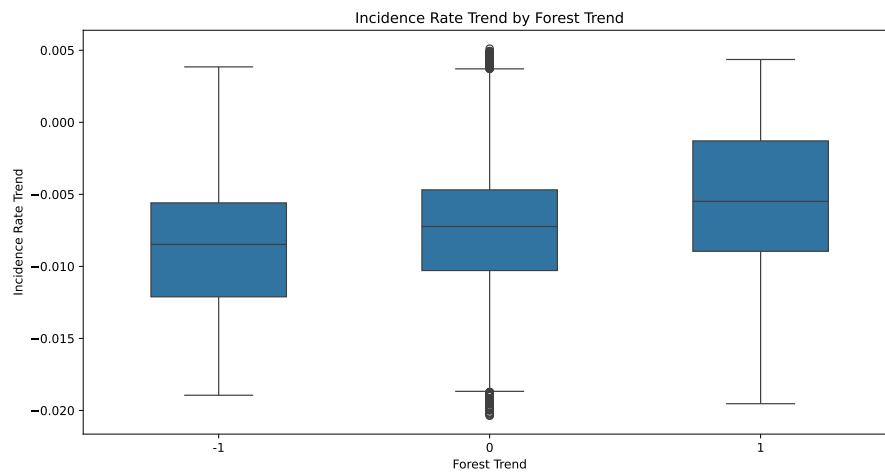
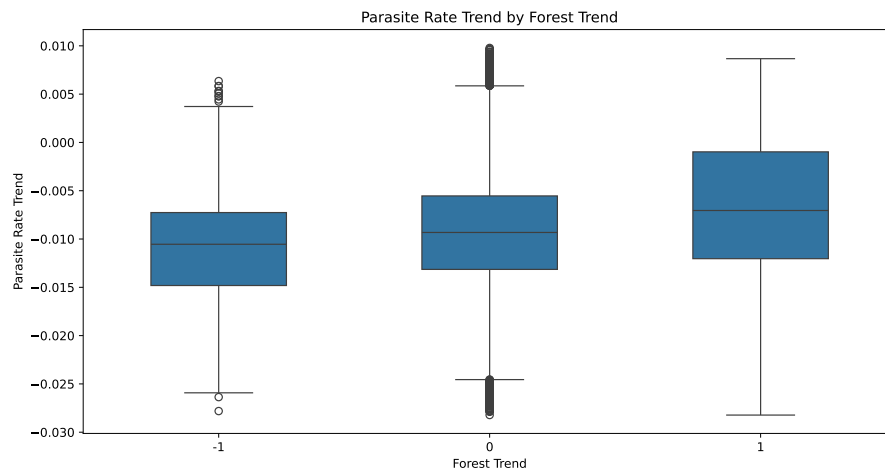
Per quanto riguarda l'analisi con immagini satellitari e modelli di regressione, i risultati ottenuti sono mostrati in Tabella 8.1. La media del *Parasite Rate*, variabile target, per l'intera nazione è 0.3692 con una deviazione standard di 0.1681 e un minimo di 0.0195 e un massimo di 0.9131. Per i villaggi, la media è 0.3177 con deviazione standard di 0.1787 e un minimo di 0.0233 e un massimo di 0.8888.

(a) Incidence Rate.

(b) Parasite Rate.

Figura 8.1: Box plot degli indici malarici per categoria di forest trend.

| **Modello** | **RMSE Nazionale** | **RMSE Villaggi** |
|---|---|---|
| Decision Tree | 0.222935 | 0.178770 |
| Random Forests | 0.158992 | 0.129987 |

Tabella 8.1: RMSE medio (10-fold CV) per i modelli Decision Tree e Random Forest.

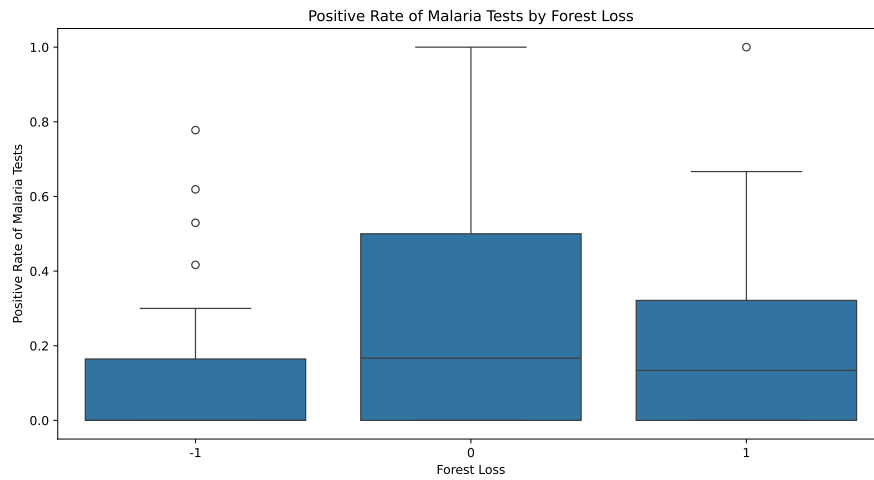



Figura 8.2: Box plot della malaria per categoria di forest loss nei villaggi.

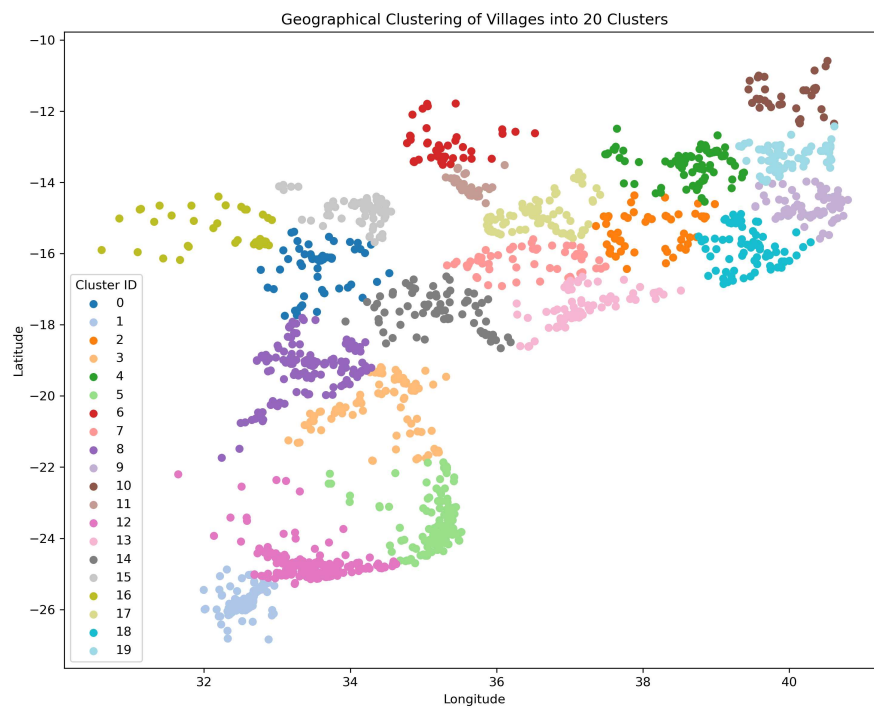



Figura 8.3: Mappa dei villaggi divisi in 20 cluster.

| Cluster | Decision Tree RMSE | Random Forest RMSE |
|---|---|---|
| 6.0 | 0.110360 | 0.084887 |
| 17.0 | 0.153209 | 0.111067 |
| 11.0 | 0.114785 | 0.086101 |
| 4.0 | 0.143551 | 0.104564 |
| 19.0 | 0.132107 | 0.098259 |
| 18.0 | 0.152715 | 0.114120 |
| 2.0 | 0.193312 | 0.139926 |
| 9.0 | 0.140185 | 0.101609 |
| 13.0 | 0.190412 | 0.146026 |
| 7.0 | 0.220187 | 0.161885 |
| 14.0 | 0.159184 | 0.118149 |
| 0.0 | 0.123299 | 0.093954 |
| 15.0 | 0.192752 | 0.142939 |
| 16.0 | 0.136073 | 0.100074 |
| 8.0 | 0.135795 | 0.099277 |
| 3.0 | 0.127070 | 0.094488 |
| 5.0 | 0.114795 | 0.085523 |
| 12.0 | 0.194054 | 0.143962 |
| 1.0 | 0.076740 | 0.059023 |
| 10.0 | 0.175041 | 0.129916 |

Tabella 8.2: Risultati medi RMSE per cluster (10-fold CV).

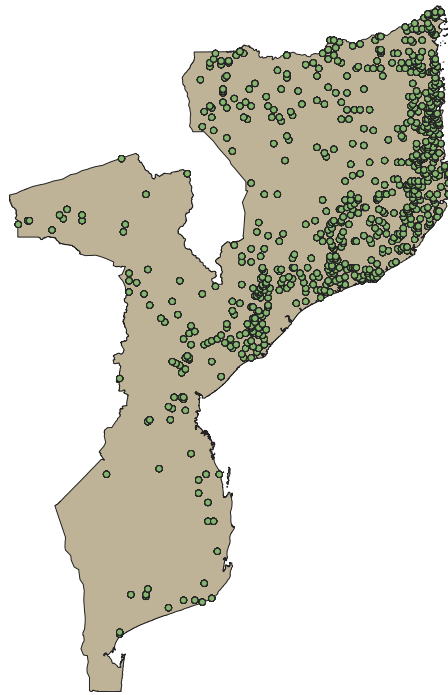



Figura 8.4: Punti filtrati in base alla deviazione standard dell'NDVI.

Dividendo i villaggi in cluster, mostrati nella mappa in Figura 8.3, i risultati del RMSE medio per ogni cluster sono presentati in Tabella 8.2.

Filtrando i punti in base alla banda *NDVI*, che ha media 0.2483 e deviazione standard 0.0731 con un minimo di -0.3393 e un massimo di 0.5002, si ottengono 12486 punti su 763856 totali e sono mostrati in figura 8.4. Ora la media dell'NDVI è 0.2654 con deviazione standard di 0.1178 quindi si è selezionato un campione con maggiore variazione di vegetazione. La matrice di correlazione relativa è rappresentata in Figura 8.5.

Le metriche RMSE per questi punti filtrati sono mostrate in Tabella 8.3. La media del *Parasite Rate* per i punti filtrati è 0.4736 con deviazione standard 0.1609. I villaggi filtrati sono 45 su 2023 totali e la media del *Parasite Rate* per questi è 0.4421 con deviazione standard 0.1509.

Vengono divisi in cluster come mostrato in Figura 8.6 e i risultati del RMSE medio per ogni cluster sono mostrati in Tabella 8.4.

Si nota, inoltre, come atteso dalla teoria, che il modello *Random Forests* ha prestazioni migliori in tutti i casi impiegando però più tempo per l'addestramento.

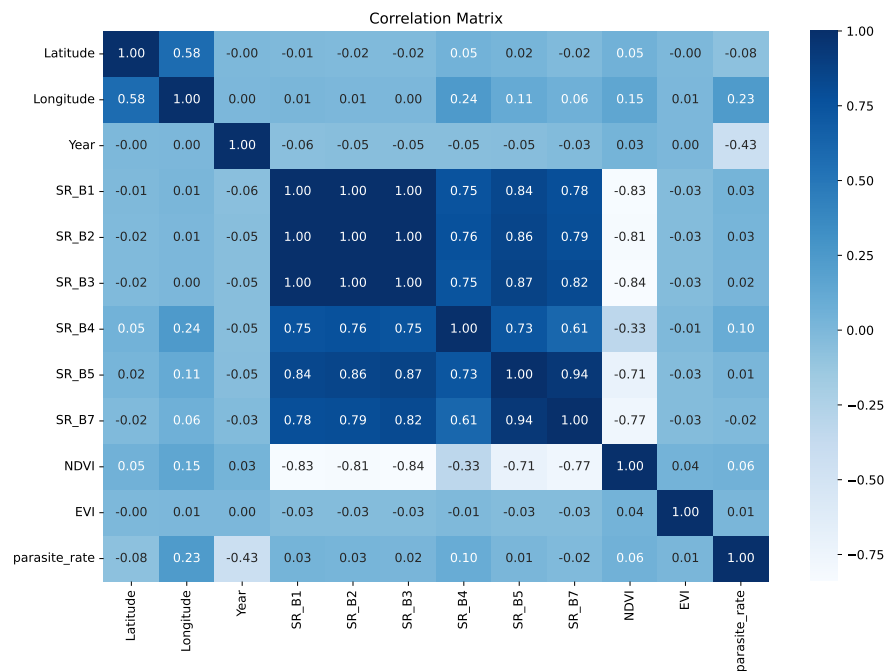



Figura 8.5: Matrice di correlazione del dataset MOZ_Sat_Mal filtrato per NDVI.

| **Modello** | **RMSE Nazionale** | **RMSE Villaggi** |
|---|---|---|
| Decision Tree | 0.222643 | 0.195468 |
| Random Forests | 0.161711 | 0.143442 |

Tabella 8.3: RMSE medio (10-fold CV) per i modelli Decision Tree e Random Forest per i punti filtrati per NDVI.

| **Cluster** | **Decision Tree RMSE** | **Random Forest RMSE** |
|---|---|---|
| 0 | 0.166140 | 0.121575 |
| 3 | 0.152549 | 0.118176 |
| 2 | 0.154937 | 0.132772 |
| 4 | 0.143037 | 0.140992 |
| 1 | 0.160444 | 0.131345 |

Tabella 8.4: Risultati medi RMSE per cluster (10-fold CV) per i punti filtrati per NDVI.

Figura 8.6: Mappa dei villaggi filtrati divisi in 5 cluster.

## 8.3 Risposta

Dall'analisi condotta, i risultati sembrano indicare l'opposto di quanto atteso inizialmente. Nel grafici sull'intera nazione, Figura 8.1, quando la copertura forestale è in diminuzione (categoria -1), anche gli indici malarici diminuiscono (coefficienti negativi) e quando la copertura forestale è in aumento (categoria 1), gli indici malarici continuano a diminuire ma in misura minore. Questo suggerisce quindi una correlazione negativa tra diminuzione della vegetazione e aumento delle infezioni malariche. Per approfondire questa relazione, oltre alle categorie LC e ai dati MAP, è stata fatta l'analisi anche nei punti dei villaggi facendo uso delle categorie LC e dei dati sui sondaggi. In questa analisi, Figura 8.2, le osservazioni erano in numero limitato e non è stata trovata una correlazione significativa risultando, anzi, contraria a quanto atteso, confermando i risultati precedenti. Questi risultati potrebbero essere coerenti con quanto osservato nel Capitolo 5, dove si è evidenziata una diminuzione sia della copertura forestale sia dei casi di malaria nel periodo considerato.

Le metriche di performance dei modelli di regressione utilizzati mostrano valori di RMSE relativamente bassi, soprattutto nei cluster più piccoli e anche dopo aver aver filtrato i punti in base alla variazione di vegetazione. Questo indica che i modelli sono in grado di predire efficacemente l'incidenza della malaria sulla base delle bande delle immagini satellitari. Questo potrebbe essere dovuto maggiormente alla relazione tra tipologia di territorio e malaria confermata precedentemente nel Capitolo 7 piuttosto che all'effettivo rapporto tra diminuzione di vegetazione e aumento malarico.

Capitolo 9

# *Conclusioni*

Dalle risposte fornite alle nostre domande di ricerca si ottiene un quadro completo, ma complesso, della relazione tra vegetazione e infezioni malariche in Mozambico. Lo studio ha evidenziato una correlazione positiva tra la presenza di vegetazione e la prevalenza di malaria, mentre non è ancora chiara la relazione tra deforestazione e infezioni, e possiamo speculare non sia presente una correlazione positiva. Nell'analisi delle tendenze generali viene osservata una diminuzione della copertura forestale e una diminuzione dei casi di malaria nel periodo dal 2000 al 2022, in accordo con la letteratura. Lo studio, poi, ha confermato che la prevalenza della malaria è più alta nelle aree con maggiore copertura forestale, in particolare nelle foreste con copertura superiore al 15%. Inoltre, questa differenza nel tasso di positività alla malaria, tra aree urbane e aree con vegetazione, persiste nonostante gli interventi governativi abbiano ridotto la prevalenza della malattia in quasi tutte le tipologie di terreno nel tempo. L'analisi delle immagini satellitari grezze, in aggiunta, ha rivelato una correlazione positiva tra l'indice di vegetazione *NDVI* e il tasso di infezione malarico, e i modelli di regressione utilizzati hanno predetto efficacemente la prevalenza della malaria sulla base delle bande delle immagini satellitari.

Lo studio ha inoltre fornito risultati utili per ricerche socioeconomiche nell'Africa subsahariana. L'analisi del territorio attorno a villaggi e miniere ha permesso di caratterizzare il contesto ambientale e identificare potenziali fattori di rischio per la malaria, e i dati raccolti sulle precipitazioni e sulla perdita forestale isolata contribuiranno a studi sui conflitti e sulle implicazioni socioeconomiche dell'attività mineraria.

Le potenziali implicazioni di questo studio per la lotta alla malaria sono molteplici. I risultati potrebbero aiutare la comprensione della relazione tra vegetazione e malaria che, soprattutto in Africa, è molto complessa ed è necessaria per sviluppare strategie di controllo efficaci e sostenibili. Si evidenzia, inoltre, come l'utilizzo di immagini satellitari e modelli di apprendimento automatico può fornire strumenti utili a monitorare la diffusione della malattia e valutare l'impatto degli interventi fatti per contrastarla. È fondamentale, infine, sottolineare l'importanza della collaborazione tra diverse discipline, come l'economia e l'informatica, e l'uso di tecniche e tecnologie avanzate per affrontare questa sfida globale per la salute pubblica.

# *Bibliografia*

[1] Leone A, Dondeynaz C, Mainardi P, Giacomassi M, Carmona Moreno C, and Chen D. *A Web Based Knowledge Management Platform For Development Cooperation in the Water Sector: Aquaknow.Net.* Chemical Industry Press, Beijing (China), 2010.

[2] Sebastian Bauhoff and Jonah Busch. Does deforestation increase malaria prevalence? evidence from satellite data and health surveys. *World Development*, 127:104734, 2020.

[3] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975.

[4] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[5] Leo Breiman. *Classification and regression trees.* Routledge, 2017.

[6] P. Defourny, C. Lamarche, S. Bontemps, T. De Maet, E. Van Bogaert, I. Moreau, C. Brockmann, M. Boettcher, G. Kirches, J. Wevers, M. Santoro, F. Ramoino, and O. Arino. *Land Cover Climate Change Initiative - Product User Guide v2*, 2017. Issue 2.0.

[7] Johannes J Feddema, Keith W Oleson, Gordon B Bonan, Linda O Mearns, Lawrence E Buja, Gerald A Meehl, and Warren M Washington. The

importance of land-cover change in simulating future climates. *Science*, 310(5754):1674–1678, 2005.

[8] Janine Felden, Lars Möller, Uwe Schindler, Robert Huber, Stefanie Schumacher, Roland Koppe, Michael Diepenbroek, and Frank Oliver Glöckner. PANGAEA – Data Publisher for Earth & Environmental Science. *Scientific Data*, 10(1):347, 2023.

[9] Peter Flach. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. Cambridge University Press, Cambridge, UK, 2012.

[10] Herbert M Gilles and David A Warrell. *Bruce-Chwatt's essential malariology*. Number Ed. 3. 1993.

[11] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017. Big Remotely Sensed Data: tools, applications and experiences.

[12] CA Guerra, RW Snow, and SI Hay. A global assessment of closed forests, deforestation and malaria risk. *Annals of tropical medicine and parasitology*, 100(3):189, 2006.

[13] Carlos A Guerra, Simon I Hay, Lorena S Lucioparedes, Priscilla W Gikandi, Andrew J Tatem, Abdisalan M Noor, and Robert W Snow. Assembling a global database of malaria parasite prevalence for the malaria atlas project. *Malaria journal*, 6:1–13, 2007.

[14] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013.

[15] Simon I Hay, Carlos A Guerra, Andrew J Tatem, Peter M Atkinson, and Robert W Snow. Urbanization, malaria transmission and disease burden in africa. *Nature Reviews Microbiology*, 3(1):81–90, 2005.

[16] Simon I Hay, Carlos A Guerra, Andrew J Tatem, Abdisalan M Noor, and Robert W Snow. The global distribution and population at risk of malaria: past, present, and future. *The Lancet infectious diseases*, 4(6):327–336, 2004.

[17] Simon I Hay, David J Rogers, Jonathan F Toomer, and Robert W Snow. Annual plasmodium falciparum entomological inoculation rates (eir) across africa: literature survey, internet access and review. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94(2):113–127, 2000.

[18] Simon I Hay and Robert W Snow. The malaria atlas project: Developing global maps of malaria risk. *PLOS Medicine*, 3(12):1–5, 12 2006.

[19] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.* The MIT Press, Cambridge, Massachusetts, USA, 2015.

[20] DA King, Catherine Peckham, JK Waage, Joe Brownlie, and Mark EJ Woolhouse. Infectious diseases: preparing for the future, 2006.

[21] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[22] Estrella Lucena-Sánchez, Guido Sciavicco, and Ionel Eduard Stan. Feature and language selection in temporal symbolic regression for interpretable air quality modelling. *Algorithms*, 14(3), 2021.

[23] Songrit Maneewongvatana and David M Mount. Analysis of approximate nearest neighbor searching with clustered point sets. *arXiv preprint cs/9901013*, 1999.

[24] Federico Manzella, Giovanni Pagliarini, Guido Sciavicco, and Ionel Eduard Stan. The voice of covid-19: Breath and cough recording classification with temporal decision trees and random forests. *Artificial Intelligence in Medicine*, 137:102486, 2023.

[25] D Metselaar and PH Van Thiel. Classification of malaria. 1959.

[26] Tom M Mitchell and Tom M Mitchell. *Machine learning*, volume 1. McGraw-hill New York, 1997.

[27] Malla Reddy College of Engineering & Technology. Introduction to data science (r20ds501) - lecture notes. https://mrcet.com/downloads/digital_notes/CSE/III%20Year/AIML/Introduction%20to%20Datascience.pdf, 2023. B.Tech III Year – I Sem (R20) – Department of Computational Intelligence, Malla Reddy College of Engineering & Technology, Secunderabad, Telangana, India. Accessed: 2024-11-16.

[28] Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[30] Daniel A Pfeffer, Timothy CD Lucas, Daniel May, Joseph Harris, Jennifer Rozier, Katherine A Twohig, Ursula Dalrymple, Carlos A Guerra, Catherine L Moyes, Mike Thorn, et al. malariaatlas: an r interface to global malariometric data hosted by the malaria atlas project. *Malaria journal*, 17:1–10, 2018.

[31] Shi Qiu, Zhe Zhu, Rong Shang, and Christopher J. Crawford. Can landsat 7 preserve its science capability with a drifting orbit? *Science of Remote Sensing*, 4:100026, 2021.

[32] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[33] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[34] Niles Ritter and Mike Ruth. The geotiff data interchange standard for raster geographic images. *International Journal of Remote Sensing*, 18(7):1637–1647, 1997.

[35] Niles Ritter, Mike Ruth, Brett Borup Grissom, George Galang, John Haller, Gary Stephenson, Steve Covington, Tim Nagy, Jamie Moyers, Jim Stickley,

et al. Geotiff format specification geotiff revision 1.0. *SPOT Image Corp*, 1:154–172, 2000.

[36] Guido Sciavicco and Stan Ionel Eduard. Knowledge extraction with interval temporal logic decision trees. *arXiv preprint arXiv:2305.16864*, 2023.

[37] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[38] David L Smith and F Ellis McKenzie. Statics and dynamics of malaria infection in anopheles mosquitoes. *Malaria journal*, 3:1–14, 2004.

[39] David L. Smith, F. Ellis McKenzie, Robert W. Snow, and Simon I. Hay. Revisiting the basic reproductive number for malaria and its implications for malaria control. *PLoS Biol*, 5(3):e42, 2007.

[40] DL Smith, J Dushoff, RW Snow, and SI Hay. The entomological inoculation rate and plasmodium falciparum infection in african children. *Nature*, 438(7067):492–495, 2005.

[41] Robert W Snow and Kevin Marsh. The consequences of reducing transmission of plasmodium falciparum in africa. 2002.

[42] Robert W Snow, Judy A Omumbo, Brett Lowe, Catherine S Molyneux, Jacktone-O Obiero, Ayo Palmer, Martin W Weber, Margaret Pinder, Bernard Nahlen, Charles Obonyo, et al. Relation between severe malaria morbidity in children and level of plasmodium falciparum transmission in africa. *The lancet*, 349(9066):1650–1654, 1997.

[43] Mark Swanson, Jerry Franklin, Robert Beschta, Charlie Crisafulli, Dominick Dellasala, Richard Hutto, David Lindenmayer, and Frederick Swanson. The forgotten stage of forest succession: Early-successional ecosystems on forest sites. *Frontiers in Ecology and The Environment - FRONT ECOL ENVIRON*, 9, 03 2010.

[44] John Wilder Tukey. *Exploratory data analysis: Limited preliminary Ed.* Addison-Wesley Publishing Company, 1970.

[45] U.S. Geological Survey. *Landsat 7 Data Users Handbook*, version 3.0 edition, 2024. LSDS-1927, Accessed: 2024-10-31.

[46] Ian H Witten and Eibe Frank. Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77, 2002.